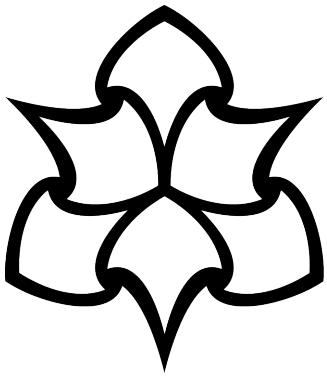


MACHINE LEARNING FOR UNDERSTANDING COMPLEX, INTERLINKED SOCIAL DATA

CLAIRE LITTLE

A thesis submitted in partial fulfilment of the requirements of the Manchester Metropolitan University for the degree of Doctor of Philosophy

Centre for Policy Modelling
Faculty of Business and Law
Manchester Metropolitan University
April 2018

ABSTRACT

With the growing availability of 'big' data, increasing computer power, and improved data storage capacities, machine learning techniques are now frequently employed in order to make sense of data. Yet, the social sciences have been slow to adopt these techniques, and there is little evidence of their use in some academic fields. This thesis explores the methods most commonly utilised in social science research, that is, linear regression and null hypothesis significance testing, in order to identify how machine learning methods might complement these more established methods.

A case study exploring the Troubled Families programme provides a practical example of how machine learning techniques can be utilised on complex, interlinked social data in order to provide deeper understanding and more insight into the data. Eleven different types of families were identified using cluster analysis, and analysis was performed in order to understand how the family's lives changed after joining the TF programme when compared to before. The analysis provided insight into the various types of families that existed and the problems that they had. It also highlighted that, had the data been analysed on an overall global level, it would have been prone to an averaging effect whereby many of the changes that occurred were not apparent; analysis on the cluster-level resulted in identification of cluster-level patterns, and a greater understanding of the data.

This thesis demonstrated that machine learning techniques, such as cluster analysis and decision tree learning, can be effectively utilised on complex 'real-life' social science datasets. These methods can identify hidden groups and relationships, and important predictors in a dataset, provide a better understanding of the structure of the data, and aid in generating research questions and hypotheses.

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Professor Bruce Edmonds and Dr Keeley Crockett, for all their advice throughout. A special thank you to Bruce whose support has been much appreciated.

Thank you too, to everyone at the Centre for Policy Modelling over the years. And also, to the English City Council who allowed me to use the data for this research. It provided a great, and very interesting, case study.

Finally, thank you to my family for all their support

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TERMINOLOGY

Attribute	Equivalent to variables or features, e.g. a person's race or gender
Predictor	Independent variable
Target	Dependent variable
Model fit	The goodness of fit of a model describes how well it fits the data. A well fitted model predicts very close or equal to the actual target value
Overfitting	Occurs where a model learns the data too well (i.e. it captures noise as well as any underlying relationship) and cannot generalise to new data
Underfitting	Occurs where a model cannot capture any pattern in the data
Type I error	Incorrectly rejecting a true null hypothesis (false positive)
Type II error	Incorrectly accepting a false null hypothesis (false negative)
Sensitivity	Also known as true positive rate or recall, this is the proportion of positives correctly identified in a classification model (e.g. the proportion of patients correctly identified as having a disease)
Specificity	Also known as the true negative rate, this is the proportion of negatives correctly identified in a classification model (e.g. the percentage of healthy people correctly identified as not having a disease)
P-value	The probability of obtaining this (or more extreme) data, given that the null hypothesis is true
NHST	Null hypothesis significance testing
TF	Troubled Family
ECC	The English City Council who provided the TF data for this study (and wished to remain anonymous)
CIN	Children in Need
CPP	Child Protection Plan
LAC	Looked After Children
NEET	Not in Education, Employment or Training

1 INTRODUCTION

1.1 BACKGROUND

The analysis and understanding of data is fundamentally important to society. Analysis of data might be used to: test an existing hypothesis; formulate new hypotheses; check the reliability and quality of a dataset; evaluate the level of impact of one or more variables against another; gain insights and explore hidden relationships; show that a dataset is not completely random; or to predict future cases or events. There are many reasons for data analysis, but perhaps the overall aims are to discover something useful, to predict something, or to support decision and policy making.

In recent years data-driven analysis (or data science) has become more prevalent, both in academic research and in industry. This has coincided with the growth of 'big data', a reduction in the cost of data storage, and ever more powerful computers. Data is being stored at a rapid rate and in massive volumes, and it is collected from almost every imaginable source - for example, customer databases, social media, sensor data, financial data, governmental data, and biomedical and scientific data. This ever-increasing store of data has generated great interest into how best to derive insights or gain competitive advantage from it. Many traditional statistical techniques are simply not equipped to cope with the size, type and dimensionality of these large quantities of data. This has prompted improvements in the methods used to analyse it, such as more efficient algorithms and the development of free, open-source software.

Whilst much of the focus has been on 'big' data with regards to machine learning, the methods can usefully be deployed upon 'smaller' data, and they are particularly suited to the types of data that are synonymous with social science research (such as social surveys, and wide, interlinked data). Academic fields such as computer science have deep involvement in developing methods for, and analysing, this data; however, the social sciences have seemingly lagged behind in adopting these new methods. Social scientists are ideally placed to ask informed research questions of this new data, to aid in understanding results and to take advantage of these methods. Yet, whilst there is growing interest in the use of machine learning methods, they are not frequently utilised in social science research.

1.2 AIMS

This project aims to show that machine learning techniques can be utilised effectively on social science datasets and that they can be particularly useful for identifying patterns and hidden underlying relationships.

The project also aims to show that machine learning techniques can effectively complement the more established methods, such as regression. Machine learning might be used for exploratory data analysis, to discover hidden groups within data, to identify relationships and important predictors, determine the structure of a dataset, to generate hypotheses, or be utilised to confirm results. The adoption of machine learning methods such as cross-validation could also produce more robust work when applied to established methods.

1.3 RESEARCH QUESTIONS AND OBJECTIVES

The overall research questions are:

- Can machine learning techniques be effectively utilised or adapted to facilitate the analysis and comprehension of large social science data sets?
- Which data mining methods are most effective for discovering otherwise hidden patterns within complex and often noisy social data?
- Can data mining methods provide a detailed picture of trends and patterns within a dataset?
- Can machine learning methods be utilised to suggest new hypotheses and research questions?

The objectives of this research are to explore the use of machine learning in the social sciences, and to discover how these methods might be best utilised on social science data. This includes considering the methods that are currently employed and exploring why there may be a reluctance to explore the utilisation of more data-driven methods.

A practical analysis of the use of machine learning methods will be provided by a large analysis of data pertaining to the Troubled Families Programme. The data is anonymised so as not to identify individuals or families. It is also noisy and interlinked, and contains information pertaining to the events that families had (such as school absence, criminal offences and child safeguarding events) and details that described the families and the

individuals in them (such as age, location, and details of the intervention treatments that families received).

This analysis will consider whether machine learning methods can be effectively utilised to identify groups or patterns within the data. More explicitly, it asks:

- Whether there exist unique groups of families within the Troubled Families data
- Does the identification of these groups provide deeper insight than one overall analysis of the data might provide?
- How do the lives of the families in each cluster change following their introduction to the TF programme, and is it possible to predict, or identify important factors that may indicate where positive future outcomes will occur?

1.4 OUTLINE

Chapters 2 and 3 provide a description of the methods that are most commonly used in social science research, this is in order to provide a baseline for the following chapters which discuss alternative methods that might be used to complement these methods. Chapter 2 provides a description of linear regression, and the various assumptions that must be satisfied in order for results to be valid. It explores model interpretation and considers the issues with linear regression that might lead to flawed results.

Chapter 3 considers null hypothesis significance testing and explores the various issues surrounding its use, a broader discussion of reproducibility is also included. Chapter 4 provides an outline of data mining, including a brief history, and outline of various methods. The methods utilised in this thesis are explored; these are clustering, decision tree learning, random forests and boosted methods. There is a discussion of both the positive and negative aspects of data mining. Chapter 5 explores the use of data mining methods in existing social science research and considers the ways that these methods might be optimally utilised.

Chapters 6 and 7 comprise parts 1 and 2 of the case study. This explored data surrounding the Troubled Families programme, which was set up by the UK Government in order to target and provide help to families with multiple problems (such as school exclusion, child safeguarding issues, or criminal offences). Chapter 6 provided a description of the Troubled Families Programme and aimed to discover whether there were any clusters, or groups of similar families within the data. The geographical location

of the families was also explored in order to determine whether location might be an important factor in a family's problems. Chapter 7 continued the analysis from Chapter 6 and utilised the clusters to investigate the lives of the families in the year after they joined the TF programme. This considered whether families had shown any improvement in their circumstances one year later. Both of the chapters utilised data mining methods, such as cluster analysis, decision tree learning and visualisation techniques in order to explore and analyse the data.

Chapter 8 concludes the thesis and provides a summary of the research together with a consideration of the contributions of the research and avenues for future work.

2 REGRESSION

2.1 INTRODUCTION

The purpose of this, and the following, chapter is to discuss established approaches used in the social sciences, and to highlight the advantages and disadvantages of these methods. These methods are explored in order to establish a reason for considering the use of alternative methods, and to provide a baseline for discussion of these alternatives.

Linear regression and other correlational methods are widely used in the social sciences, and can be powerful tools, in that their results are relatively easy to understand, and they can intuitively highlight relationships between attributes. However, linear regression relies upon strict statistical assumptions to be effective and this chapter explores those assumptions and the consequences of not satisfying them.

This chapter also provides a brief description of Ordinary Least Squares Linear Regression. Methods of interpretation and evaluation of regression models are described, highlighting that there can be weaknesses with some of the methods used. The importance of visualisation and exploratory data analysis in the modelling process is considered. A discussion of the literature surrounding the various issues associated with the use of linear regression in the social sciences is provided. This explores what the consequences of misspecified regression models are, gives a consideration of why models are sometimes misspecified, and considers suggestions as to how to overcome some of the problems.

2.2 BACKGROUND

Regression Analysis is one of the fundamental processes of modern statistics and encompasses a broad range of techniques and methods, but in essence, it explores the relationship between a target (or dependent, outcome or response) attribute and one or more predictor (or independent or explanatory) attributes. It is used to identify whether there is a relationship between the predictor/s and the target attribute, to describe both the form and strength of those relationships, and to also provide an equation (or mathematical model) describing the relationship such that the predictors can be used to predict the target.

Ordinary Least Squares (OLS) Linear Regression is one of the most commonly used methods in Social Science papers that require some form of quantitative analysis of the relationships between attributes. The method had its beginnings in the early 1800s when mathematicians Adrien-Marie Legendre (1752-1833) and Carl Friedrich Gauss (1777-1855) both, independently of each other and for the purpose of calculating astronomical orbits, published papers describing the method of least squares (Sorenson, 1970).

Building upon this, scientist Sir Francis Galton (1822-1911) was instrumental in the development of modern ideas of regression and correlation. He was particularly interested in genetics and heredity, and presented the first regression line at a lecture in 1877 (Stanton, 2001). In further work on the heights of parents and their children, he first described the phenomenon of regression towards the mean (Galton, 1886). Galton's colleague, mathematician Karl Pearson (1857-1936), and Udny Yule (1871-1951) extended and formalised mathematically much of their ideas on regression and correlation (Yule, 1897; Pearson et al., 1903). Pearson himself is generally credited as being one of the founders of modern statistics, having defined the Pearson Product Moment correlation coefficient, the Chi-squared distribution, and the idea of p-values and statistical hypothesis testing (Pearson, 1900).

Although in recent years new regression methods have been developed to deal with more complex data and problems, OLS regression is still arguably the most commonly used method (Berk et al., 2014). Newer methods include: robust regression, which attempts to deal with outliers and heteroscedasticity; nonparametric regression which can provide a more flexible regression curve and relaxes the assumption of linearity; Bayesian regression which is useful for poorly distributed and more complex data and provides analysis within the context of Bayesian inference; ridge regression which attempts to alleviate multi-collinearity of predictors; and time series regression.

2.3 OLS LINEAR REGRESSION

Simple linear regression is the method of predicting a quantitative target attribute Y from a single predictor attribute X, assuming that there is a linear relationship between the two (i.e. that their relationship can be described by a straight line when plotted). Mathematically, where there are n data points, a target attribute y_i, and the single predictor x_i the relationship is written as:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (2.1)$$

For $i = 1, \dots, n$ and where the regression coefficients β_0 and β_1 are unknown constants that represent the intercept and slope terms in the model. ϵ_i is the error term or residual. The intercept β_0 is the value of y when x = 0, i.e. it is the point at which the regression line crosses the y-axis when plotted. The slope, β_1, refers to how much change there is predicted in y for one unit change in x. The residual is the difference between the predicted value ($\hat{y}_i$) and actual value of y_i, written $\epsilon_i = y_i - \hat{y}_i$. Residuals may be positive or negative; if the residuals were all zero, all data points would sit on the regression line and there would be no error at all in the model.

In many cases, a simple model is not adequate and there is a need to consider more than one predictor; this is particularly true when dealing with complex datasets and real-world problems. In this case, multiple linear regression is required. The equation for this is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \epsilon_i \quad (2.2)$$

For $i = 1, \dots, n$ and where there are n data points, y_i is the target, and x_{i1} to x_{ij} are the predictors. β_0 is the intercept and β_1 to β_j are the regression coefficients.

For both multiple and simple linear regression, the optimal solution is found by using the data to estimate the values of the unknown constants β_0 to β_j. There are various methods for accomplishing this, but by far the most commonly used (Hayes and Cai, 2007; James et al., 2013) is the method of Ordinary Least Squares, or OLS. This may be because it is relatively quick and easy to implement (it is included in most statistical software as standard), the basic methodology is relatively easy to understand (especially for those without a mathematical background), and it is easily interpretable. OLS seeks to find the straight line or plane that cuts through the data points, producing the least amount of error.

The OLS function derives the estimates ($\hat{\beta}_0$ to $\hat{\beta}_j$) by finding the line or plane that minimizes the sum of squared errors (SSE) between the actual and predicted values for y_i. Squaring the errors removes any issue over whether the error is negative or positive.

$$SSE = \sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (2.3)$$

Minimising this gives, in the case of simple regression, the following equations (James et al., 2013):

$$\hat{\beta}^1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.4}$$

$$\hat{\beta}^0 = \bar{y} - \widehat{\beta_1}\bar{x} \tag{2.5}$$

Where $\bar{y}$ and $\bar{x}$ are the means of the target and predictor attributes.

Multiple linear regression can be generalised to handle categorical target attributes and those that are not Normally distributed. In the case of a binary target, logistic regression may be used. Logistic regression aims to estimate the probability of a specific level of the target attribute given the predictors, and can be employed for both binary or multinomial target data. Linear and logistic regression (and Poisson regression, ANOVA, etc.) belong to a broader class of models called the generalised linear model, or GLM (Nelder and Wedderburn, 1972). GLM is a flexible generalisation of linear regression that allows for target attributes that follow any distribution from the Exponential family (rather than just those that are Normally distributed) and allows for a link function of the mean to vary linearly with the predictors (rather than assuming that the target itself must vary linearly).

2.4 ASSUMPTIONS

Linear regression remains popular in the social sciences because it provides a relatively easy to understand equation that can help to identify significant predictors. It also allows a researcher the ability to examine the effect of one predictor on the target whilst holding all other variables constant. This is particularly useful in a social science context, where unlike with more natural sciences (for example, chemistry), the experimental variables often cannot be manipulated. For example, a researcher cannot change a subject's income or age whilst holding all other variables constant; much social science data is observed, and not manipulated through experiments. However, as with many statistical methods, for OLS regression to be effectively implemented, and for its results to be reliable it must satisfy certain assumptions (Boslaugh, 2013; James et al., 2013):

- Linearity – The target should be linearly related to the predictor/s. That is, the relationship could be plotted on a straight line (or on an n-dimensional plane where there are n predictors)
- Normality – all continuous attributes should be approximately Normally distributed, and without extreme outliers
- Independence of errors – the prediction error for each data point should be independent of the prediction error of all other data points, and errors should be Normally distributed

- Homoscedasticity – the variance of the target should not vary for different values of the predictors. That is, the prediction errors should be constant over the entire data range, and not, for instance, smaller or larger when Y is small
- Multicollinearity – for multiple regression, none of the predictors should be correlated with each other. Highly correlated predictors can obscure the true relationship of each individual predictor to the target attribute

Linear regression also generally relies upon the idea that the data is of good quality (no measurement errors), is not too small, and where it is a sample, that it is drawn from a random sample of the whole population. In reality, when dealing with complex social science data, these assumptions can be very difficult to identify and satisfy. Often, they may be only partially satisfied, if at all. Yet if they are not, then it is likely that regression results will be inaccurate, and that conclusions drawn from the model may be misleading (Freedman, 1995; Berk et al., 2017). The severity of the consequences varies greatly depending upon the assumptions that are not satisfied. Outliers and non-linearity can cause bias in the regression parameters, meaning that the relationships are not described accurately. Heteroscedasticity, multicollinearity and residuals that are not Normally distributed result in biased standard errors of regression estimates, which then lead to incorrect confidence intervals and significance tests (Erceg-Hurn and Mirosevich, 2008). This can cause problems when making statistical inferences and generalising to a larger population.

There are methods to deal with some of these issues. For example, data might be transformed into a more linear form (e.g. using a log or inverse transformation), interaction terms might be included, or outliers removed, so that assumptions are met. But these transformations must be deemed suitable for the particular model and also be interpretable. However, often the first difficulty may simply be identifying that there is a problem, particularly when dealing with large, complex datasets. In reality, models may rarely completely satisfy all of the assumptions. Where assumptions cannot be adequately satisfied, or the data simply is not linear, it may be that a more robust or non-parametric method would be more suitable, rather than performing linear regression.

2.5 Model Interpretation

Ideally models should be interpreted and evaluated by a mixture of visualisation and statistical measures (Achen, 2005; Draper and Smith, 2014; Woodside, 2016), these are considered in the following section.

2.5.1 Visualisation

Whilst there are many statistical measures available to evaluate a model, visualisation is also a vital tool. Visualisation of data is important both before and after a regression as it may reveal problems that are not easily identifiable from numerical statistics alone. Prior to regression, visualisation may highlight when a dataset does not satisfy regression assumptions and enable suitable transformations to be made to the data where appropriate; it thus may help avoid making Type I and Type II errors. It can be particularly useful in identifying outliers, non-linear relationships, collinearity of predictors, and homoscedasticity.

After regression, visualisation can be a useful tool for analysing the residuals and may also help to identify possible weaknesses in the model, such as heteroscedasticity. Useful methods include plotting the residuals against the fitted values and predictors, and quantile (Q-Q) plots to check for normality.

Given how useful it can be, visualisation should be a vital step in the regression process, yet in general, statistical visualisation is frequently underused in the social sciences (Zuur et al., 2010; Healy and Moody, 2014). Despite the increased availability and usability of software to aid in the production of visualisations (and in data analysis generally), it seems that the social sciences, and in particular sociology, are lagging behind other areas in their use of statistical visualisation (Healy and Moody, 2014). As highlighted by Healy and Moody (2014) it is common for top sociology journals to publish papers with many tables but no figures, whereas the opposite is true of the top natural science journals; in these, a key figure is often central to the article and this may help to illuminate the discussion.

The consequences of relying solely upon descriptive statistical outputs and not adequately exploring and visualising data were effectively highlighted by Anscombe (1973) and more recently by Soyer and Hogarth (2012). Despite its age, the Anscombe's quartet example is still widely quoted (for example, Shoresh and Wong (2012), Branch (2014), Healy and Moody (2014), Lindsay (2015), Nielsen (2016) and Woodside (2016)) because it effectively illustrates that exactly the same summary and regression output statistics can be obtained from very different datasets.

Anscombe (1973) created four simple datasets of x and y values, with 11 observations for each, and performed a regression on each dataset. The data is contained in Table 1; in

the first three datasets the x values are identical, however, the y values are different for all datasets. Table 2 highlights the summary statistics for each dataset, which are almost identical despite differences in the data.

Table 1: Anscombe's quartet data, x and y values for four datasets, from Anscombe (1973)

Dataset 1		**Dataset 2**		**Dataset 3**		**Dataset 4**	
x1	**y1**	**x2**	**y2**	**x3**	**y3**	**x4**	**y4**
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Table 2: Anscombe's quartet summary statistics for all four datasets, from Anscombe (1973)

Statistics for all four datasets:	**Value:**
Mean of x	9 (exact)
Variance of x	11 (exact)
Mean of y	7.50 (to 2 d.p.)
Variance of y	4.1 (to 1 d.p.)
Correlation between x and y	0.82 (to 2 d.p.)
Linear regression equation	y = 3.0 + 0.5x (to 1 d.p.)
$\boldsymbol{R^2}$	0.67

The R^2 for all the models is the same, and is moderately high, explaining two thirds of the variation in the data. However, when the x and y values for each dataset were plotted together with the regression lines an interesting picture emerged (Figure 1). For Dataset 1, the model had captured the relationship, with the regression line going directly through the middle of the points, but for the other three models there were clear problems. Dataset 2 showed no linear relationship between x and y (so linear regression was not appropriate, at least without transformation). Dataset 3 had one clear outlier which had forced the regression line upwards (without it, there would be a straight line). And dataset 4 also had one clear outlier that had skewed the relationship from what would have been a simple straight line.

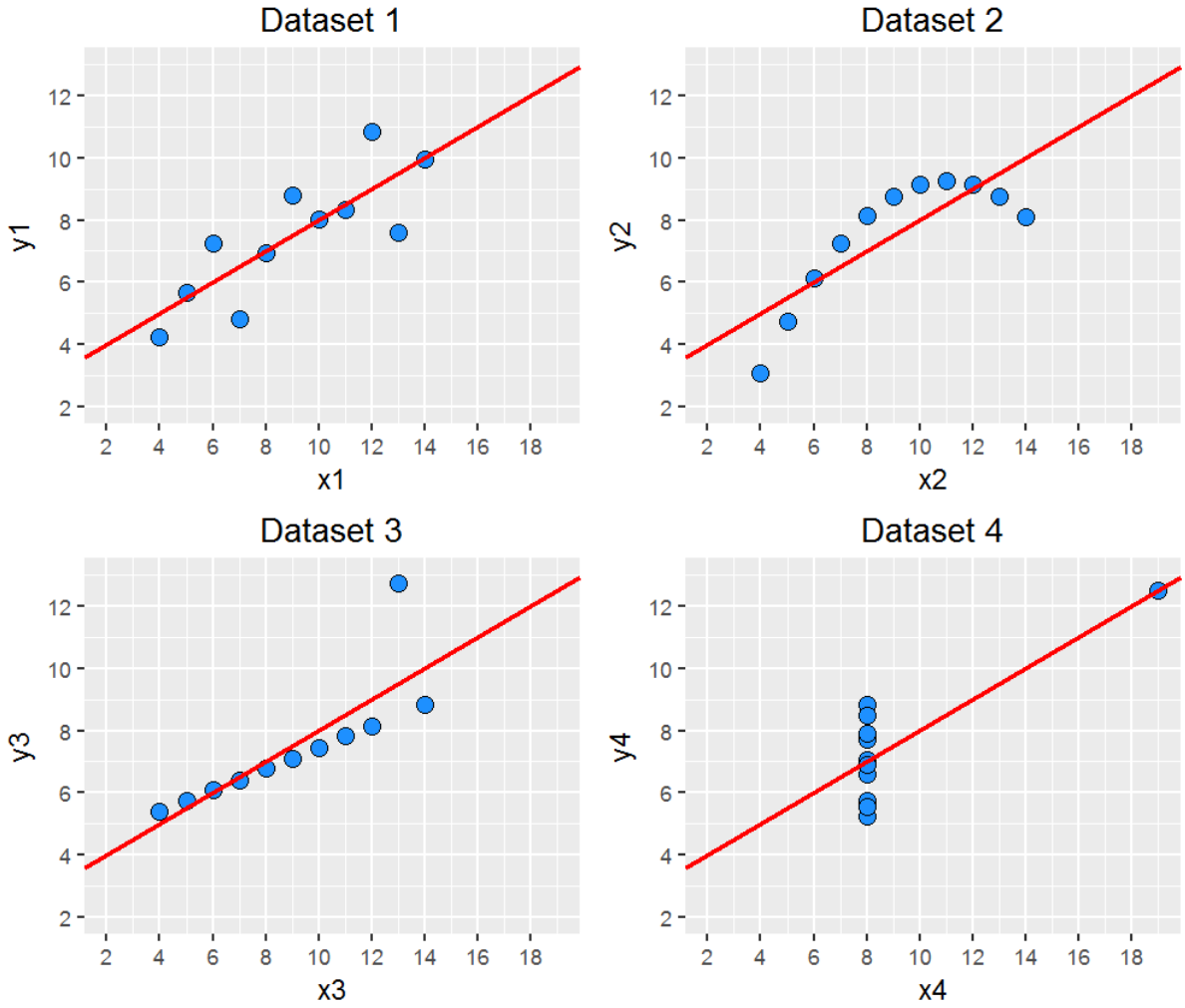


Figure 1: Anscombe's Quartet data (Anscombe, 1973) plotted to illustrate the importance of data visualisation

Whilst a simple example in two dimensions, Anscombe highlighted the importance of not relying solely upon numerical statistics, and of visualising the data both before and after a regression. It also highlighted that, in this case, evaluating the models using the R^2 value would have led to a belief that all models were equally valid, when they clearly were not. Another example is provided by Matejka and Fitzmaurice (2017) who created multiple datasets that wildly differed visually but shared identical summary statistics.

Contrasting Anscombe's work, Soyer and Hogarth (2012) illustrated that providing numerical regression outputs alone can result in misleading interpretations of regression models. Their experiment asked leading academic economists to interpret simple linear regression outputs (such as R^2, standard errors, regression coefficients and scatter plots) and make probabilistic inferences from them. The participants produced the most accurate interpretations where given only visualisations to study. When given just numerical regression outputs they were less accurate; interestingly, the addition of visualisations to these made little difference. Where forced to consider only

visualisations (scatter plots with regression lines) participants were more accurate and Ziliak (2012:713) suggests that this was because a simple graph allowed them to visualise the model uncertainty, whereas without this the econometricians fell back upon focussing on R^2 and t values and hence were likely to 'vastly' over or underestimate the levels of uncertainty.

As acknowledged by the authors, there were limitations to the study. It had only a 9% response rate, and asked questions which might be considered 'tricky' as they involved calculating probabilities (as opposed, say, to simply assessing the 'significance' of particular variables). However, it highlighted the extent to which regression outputs alone may be misinterpreted, and identified that providing basic visualisations allowed the experts to infer levels of uncertainty more accurately.

Both the Soyer and Hogarth (2012) and Anscombe (1973) examples highlight that visualisations can allow more accurate inference, and that summary statistics alone cannot always identify correlations and non-linear relationships within a dataset. However, these were both simple, two-dimensional examples and it should be noted that more complex, higher dimensional data may be more difficult to analyse. Newer data mining visualisation methods present ways to do this (Liu et al., 2017), but seem so far to be more frequently utilised outside the field of social science. In particular, as well as more established methods such as Principal Component Analysis, a method such as t-Distributed Stochastic Neighbor Embedding (t-SNE) is effective at representing high-dimensional data in two (or three) dimensions, which may then be visualised in a scatterplot (Van Der Maaten and Hinton, 2008).

2.5.2 Statistical Measures

Statistical measures used to evaluate a regression model generally include the F statistic, Residual Standard Error (RSE), coefficient of determination (R^2), the standard errors of the regression coefficients and their t-values and p-values, and tests to determine the distribution of the residuals, such as the Kolmogorov-Smirnov test (Boslaugh, 2013; James et al., 2013). Where there are multiple predictors, the standard errors and their p-values are typically used to decide which predictors are significant.

As well as considering the standard errors of the individual regression coefficients, the Residual Standard Error (RSE) is also a popular measure of overall fit (or lack of fit) of a

regression model (James et al., 2013). It is an estimate of the standard deviation of the residuals, ϵ, and has the formula

$$RSE = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2.6)$$

Where n is the sample size, p is the number of predictors, and $y_i - \hat{y}_i$ is the error, or residual.

RSE is measured in units of Y, and not in the form of a percentage or proportion. This means it should provide a meaningful measure of fit (since it is in the same unit as the target), but in practice it can be difficult to determine what a good RSE value is. However, a small RSE generally indicates that there is only small error and the model fits the data well, i.e. that $y_i \approx \hat{y}_i$ for *i = 1,…,n*. Whereas a large RSE may indicate a lack of fit, i.e. that the values of $\hat{y}_i$ are very far from y_i. It is dependent upon the specific dataset as to what might constitute acceptably 'small' or 'large' RSE values. An advantage of the RSE is that it can be used to compare different regression models (as opposed to the R^2, which cannot where different data samples are used).

2.5.2.1 R-Squared

Perhaps the most commonly used method of evaluating model fit (Renaud and Victoria-Feser, 2010; Draper and Smith, 2014) is to calculate the coefficient of determination, or R^2. The R^2 is generally defined as the proportion of variation in the target, Y, that is explained by the model. R^2 takes a value between 0 and 1, where an R^2 of zero would mean a model that explained none of the variation, and an R^2 of 1 would mean a perfect model (all points on the regression line/plane). An R^2 of 0.9, for example, would mean that 90% of the variation in the values of Y could be accounted for by the values of X.

For simple linear regression the R^2 is calculated as the square of the correlation between the target (Y) and predictor attribute (X), that is $Cor(X,Y)^2$. In the case of multiple linear regression, R^2 is calculated as the square of the correlation between the target and the prediction, $Cor(Y,\hat{Y})^2$.

Adding more predictors to a model will always increase the R^2, regardless of whether there is any significant effect (James et al., 2013:212). Therefore, it can be a deceptive metric where there are many predictors. An alternative measure is the adjusted R^2,

which corrects for sample size and the number of predictors, by penalizing the addition of predictors that add nothing to the model (James et al., 2013:212).

$$Adjusted\ R^2 = 1 - \frac{SSE/(n-p-1)}{TSS/(n-1)} \tag{2.7}$$

where p is the number of predictors, n the data sample size, SSE is the sum of squared errors, and TSS is the total sum of squares.

The adjusted R^2 may be a more suitable measure than R^2 if a model contains many predictors, or when comparing models that contain different numbers of predictors (but use the same dataset). However, despite the apparent usefulness of the adjusted R^2 in these conditions, it seems that it is still not utilised as often as it could be in academic research. This may be because some regression texts do not seem to stress its usage (for example, Draper and Smith (2014:140)) or because it is seen simply as a measure to enable choice between many models and this is often not required.

Overall, the R^2, with its value measured as a proportion between 0 and 1, can be an attractive metric as it appears much easier to understand than measures such as the RSE. However, the R^2 value can be deceptive, and in practice, it is not always easy to determine what a 'good' R^2 value should be. It is dependent upon the context of the model; the particular dataset and the model goal. In fields such as machine learning, a value close to 1 would be expected from a 'good' model, whereas in fields that contain noisy data with much unmeasured error, a much smaller R^2 may be deemed acceptable. Reporting R^2 values of less than 0.1 is common in fields such as sociology and political science and whilst such low values could indicate measurement difficulties or large random effects, it could also indicate that important factors have been omitted from the model (Freedman, 2009:52).

The R^2 has long been considered a controversial measure of fit for regression models, with the criticism spanning many decades (for example, Tufte (1969), Achen (1977), King (1991), Berk (2004)). A problem with using the R^2 as a measure of fit is that in the simple case, it does not evaluate the model at all, it is simply an indicator of correlation. In the case of multiple regression, the R^2 always increases as the number of predictors increases, irrespective of model performance, and this can be misleading. And since the R^2 is dependent upon underlying variation in the data, it cannot be used to compare models built upon different data samples, therefore two (or more) model R^2 may differ

simply because of different sample variance, rather than that the underlying relationships have changed (Achen, 1977).

Achen (1990) describes the R^2 as meaningless and a measure only of the particular data sample, making it useless for determining the quality of model fit. Whereas (King, 1986) points out that there is no statistical theory behind the R^2, it simply measures the spread of data points around the regression line/plane. Despite the criticism over the years, it seems it is not uncommon for social science academic papers to report only the R^2 value of their models when referring to overall fit. The previous example of Anscombe's (1973) quartet highlighted the danger of evaluating a model using only the R^2; all four models had the same R^2 despite three of them being completely misspecified. Correlation, in general, is a poor way of summarising data; Tufte (1969) also performed a very similar experiment to Anscombe's, which highlighted the faults of the correlation coefficient by plotting three datasets all with the same correlation, but very different data distributions.

However, despite the criticism, much of the critical literature still make the point that the R^2 can be useful. It is a useful first metric to consider (Draper and Smith, 2014:34), in the sense that, superficially at least, a high value may provide some indication that the model has captured relationships between the predictor and target attributes for that particular dataset, whereas a low value indicates that the model probably has failed to capture relationships, if there are any. As Freedman (2009:53) states 'the R^2 measures goodness of fit, not the validity of any underlying causal mechanism', and this means that other statistical metrics must also be considered when evaluating a model. The R^2 should not be reported in isolation (Luskin, 1991), as this can be misleading. For instance, a high R^2 value without any individual significant regression coefficients might warrant further investigation into whether the model meets all assumptions, such as multicollinearity.

2.6 CRITICAL LITERATURE SURROUNDING THE IMPLEMENTATION OF LINEAR REGRESSION IN THE SOCIAL SCIENCES

Regression models can be a powerful tool for social scientists, in that their results are relatively easy to understand, and they can highlight relationships between attributes. Where assumptions are well understood, regression models can provide useful explanatory power of a particular phenomenon and detail the effects of each predictor upon a target whilst holding all others constant (or controlling for). This is all without

requiring much data (they can be used with relatively small datasets) or computing power. However, criticism of the use of linear regression in social science research has been around for many years (for example, Leamer (1983), McGregor (1993), Freedman (1995), Wilcox (1998), Berk (2004), Achen (2005), Erceg-Hurn & Mirosevich (2008), Elwert & Winship (2010), Armstrong (2012), Woodside (2016)). The criticism generally focusses on the misuse of the method (whether knowingly or not), and unrealistic expectations that surround its usage. The method itself is not criticised, simply its implementation.

Perhaps the main problem associated with the use of linear regression is that it relies on very strict statistical assumptions which are extremely difficult to satisfy (Berk et al., 2017); failure to satisfy these assumptions means that the method is often misused and incorrectly applied. It seems that little attention is paid to this problem, yet if a model is misspecified and initial regression assumptions are not satisfied then this can result in Type I and Type II errors, leading to inconsistent research and the production of flawed conclusions (McGregor, 1993; Freedman, 1995; Berk et al., 2017).

Wilcox (1998) makes the point that there has been no shortage of academic research over the years stating that methods such as OLS linear regression, ANOVA and other correlational methods are not robust when the underlying assumptions are not met. And regression assumptions are rarely satisfied when analysing 'real' data (McGregor, 1993; Freedman, 1995; Berk, 2004; Erceg-Hurn and Mirosevich, 2008).

McGregor (1993:802) states that regression assumptions are 'almost always ignored, dismissed, left unexamined, or consciously violated', and argues that regression models have been an impediment to progress in the social sciences, in that ignoring assumptions can result in misleading errors and wrong conclusions. Similarly, Freedman (1995) argues that the regression models used by social scientists to make causal inferences generally depend upon many untested and unarticulated assumptions. Models built upon such a foundation are prone to a lack of reproducibility, and reliance upon their results can be misleading.

Perhaps the simplest assumption of a regression model is that the relationship between the target and predictor attributes is linear and additive. That is, that the relationship could be plotted on a straight line (or an n-dimensional plane where there are n predictors). Yet in reality, this represents such a simple model of any problem, and given the complex nature of many social science research questions, it is unlikely that many

models would truly fall into such a linear relationship. It therefore seems that the method of linear regression is often inappropriate for usage on complex social science data (McGregor, 1993).

Heteroscedasticity is a regression assumption which can be difficult to identify (Erceg-Hurn and Mirosevich, 2008); it can be caused by non-linear relationships, interactions, incorrect scaling of data or the existence of different groups within the data. Although heteroscedasticity should not bias the estimate of the regression coefficients, it can affect the validity of significance tests and confidence intervals, producing liberal or conservative estimates and therefore lead to Type I and Type II errors (Hayes and Cai, 2007).

Another regression assumption that can be particularly difficult to detect and satisfy, even for trained statisticians, are interactions. When dealing with complex, high-dimensional data it may be almost impossible to realistically identify all interactions. Even for low-dimensional data, the complexity of much social science data means that relationships within the data may not be clearly understood. Yet not detecting and accounting for interactions means that estimates are likely to be biased (Elwert and Winship, 2010). Whilst there are methods to aid in the detection of interactions, such as the use of group-level variables, these are often not employed (Erceg-Hurn and Mirosevich, 2008).

Dawson (2014) suggests that studies containing interactions are found in almost all journals containing quantitative research, yet, in general, researchers are not well equipped to either recognise or deal with them. Many research papers do not even mention whether they have tested for, or considered, the presence of interactions (Vatcheva et al., 2016). Elwert and Winship (2010:327) assert that despite the fact that most models will contain interactions of some kind, the 'overwhelming majority' of OLS regression models in the social sciences count all predictors as main effects; interactions are simply ignored. It is likely that the suggestion it is the 'overwhelming majority' of models might be an overestimation (as the authors provide no evidence to back this up), but there is little doubt that many social science regression models do not (or simply cannot) adequately account for interactions.

(Elwert and Winship, 2010) surmise that this may be because, although social scientists are aware of effect heterogeneity (i.e., they would acknowledge that causal effects may

vary from group to group, for example), they have an implicit belief that the main effects coefficients of a model provide an 'average' of causal effects. That is, social scientists simply hope that their failure to identify interactions in their model will result in returning 'average' causal effects without biasing the model. This approach has been shown to be unreliable where effect heterogeneity exists; main effects only models can be effective in some cases, but not in others (Elwert and Winship, 2010). This has implications for inference, as since heterogeneity in social phenomena is so prevalent, it is dangerous to extend results to the general population, as it cannot be clear whether results will generalise (Xie, 2013).

Not satisfying regression assumptions such as Normality means that any resulting confidence intervals and effect sizes may be inaccurate, and whilst there are robust regression techniques that can deal more effectively with outliers, skewness and non-Normality, they are rarely employed (Erceg-Hurn and Mirosevich, 2008). Wilcox (1998) argues that psychology journals contain many nonsignificant results that would have been deemed significant had more modern (post 1960) robust techniques been used. This view is reiterated by Erceg-Hurn and Mirosevich (2008) who surmise that many researchers are simply unaware that classical parametric tests such as regression have limitations, or that there exist more robust methods that might overcome this. Robust methods suggested by Erceg-Hurn and Mirosevich (2008) and Wilcox (1998) include using trimmed means, Winsorized variances, rank-based methods and bootstrapping.

Overall, there are likely a number of reasons for the problems with satisfying regression assumptions that are detailed in the literature: researchers might be unaware of the assumptions, or else simply do not have a clear understanding of them; researchers might understand but have difficulty in identifying problems or implementing a solution with complex models (Erceg-Hurn and Mirosevich, 2008); or researchers might simply ignore the assumptions (Berk, 2004). Part of the difficulty in identifying assumptions may be because people are simply not taught those skills. In general, there is a lack of analytic skills to deal with data (Peng, 2015), and many undergraduate courses teach only basic regression and do not cover more advanced methods (Eisenhauer, 2015).

It may be that much of the critical literature is simply overlooked, and that perhaps without direct practical examples it is difficult for a researcher to appreciate how the reported issues might impact upon their research. If a regression model is utilised simply

to describe a dataset and is not responsible for providing inference to a larger population, then perhaps much of the criticism seems overly negative, as much of the focus is on the consequences for inference. Berk (2004) lists several examples of regression models that successfully identified trends or patterns in data, without the need for statistical inferences or causal statements.

However, where regression models are responsible for inference and policy decisions, the consequences of their misuse can be more serious. It is notable that much of the quoted literature, whilst providing technical detail, do not provide many practical examples of the specific consequences for a regression that has been misspecified; but they may feel that the mathematical detail should suffice. One of the consequences of misspecified regression models is perhaps more broadly evident in inconsistent research results, that is research that produces conflicting conclusions, does not replicate (Open Science Collaboration, 2015) or is deemed unreliable (Berk, 2004; Ioannidis, 2005). As McGregor (1993:802) notes there is often 'no corpus of reinforcing findings' from regression studies that cover the same factors - they can result in very varied regression coefficients and model fit values, but if the regression model was strong, then similar studies should lead to similar results, however, they do not (McGregor quotes particular examples of the covariates of democracy). Ward et al. (2010) make a similar point about the predictors of civil conflict, arguing that despite various large studies being conducted, there is little accurate guidance in this area. The National Research Council (2012:1) of the USA decided that since the various studies into the deterrent effect of capital punishment had reached 'widely varying, even contradictory, conclusions', (for example, concluding that executions save lives, or that they actually increase homicides, or that they have no effect) then research studies should not be used to inform policy judgement about capital punishment. They made the point that the studies were 'plagued by model uncertainty' and the regression models used strong assumptions that lacked credibility (such as assuming homogeneity across states and years) (National Research Council, 2012:7).

Another example is the critical response to Donohue and Levitt's (2001) claim that the legalization of abortion in the USA in the 1970s resulted in a drop in crime rates nearly two decades later. There were various conflicting responses to this: Lott and Whitley (2001) suggested that legalizing abortion actually increased murder rates; Joyce (2004) did not find any meaningful association between the legalization of abortion and the drop

in crime; Foote and Goetz (2008) also found no link and pointed out mistakes in the initial analysis; whilst Reyes (2007) suggested that the drop in crime rates was due to the removal of lead from gasoline (but suggested the legalization of abortion was also an important factor). All the studies utilised regression analysis, and where practicable the same or similar data. The conflicting results highlight how difficult it can be to model very complex problems, and make suitable assumptions, using regression analysis (Berk, 2004).

Aside from problems associated with identifying and satisfying regression assumptions, another area of concern in the literature surrounds the overall evaluation of regression models. Breiman (2001b), Hill and Jones (2014) and Muchlinski et al. (2016) all make the point that the overall fit of the model appears to be of secondary importance in some research papers; as long as there are some significant p-values, and the model overall is deemed statistically significant, a high RSE or low R^2 (or any other reported measure) is not necessarily seen as any cause for concern. Indeed, many research papers simply ignore the fact that a model explained only a small amount of variance (Muchlinski et al., 2016). Yet regression models are often used to accept or reject hypotheses, and to determine the strength of relationships between the predictors and the target, therefore if the model fit overall is poor, it may indicate that some caution should be applied when evaluating how well the model really represents the data. In the absence of cross-validation, such weak results may be down to chance. Cross-validation, or the use of a holdout test dataset can allow a more accurate evaluation of the usefulness of a model (Breiman, 2001b; Ward et al., 2010; Hill and Jones, 2014; Muchlinski et al., 2016; Woodside, 2016).

Overall, despite the critical literature surrounding the use of regression, it seems that little has changed over the years; Berk et al (2014:423) make the point that much of the literature is unrebutted and 'research practice proceeds in much the same manner'. This may be because OLS regression is viewed as the standard accepted method of analysis in some fields, and this means that in some cases it is used regardless of any specification issues. Yet it makes little sense to perform an inappropriate, misspecified regression analysis simply because that is what is expected, or because some results have to be produced (Achen, 2005).

The literature contains varied suggestions as to how to overcome some of the problems, and some of these solutions, such as the use of robust methods, were discussed in the previous paragraphs. However, Armstrong (2012:693) points out that whilst solutions have been developed to deal with the various problems associated with regression analysis, these are 'often ignored in practice'. Perhaps the most obvious solution, where there are specification issues that cannot be overcome, is to use an alternative method to linear regression (McGregor, 1993; Breiman, 2001b; Achen, 2005).

Breiman (2001b) suggests that far greater attention should be paid to data analysis, that the data and the problem should be considered before any decision of which method to utilise is made; following this analysis, it may be that a method such as linear regression is suitable, or it may be that an 'algorithmic' method might be more suitable. Armstrong (2012) also advocates paying greater attention to data analysis, and suggests that more parsimonious models be utilised (using no more than three predictors) and that predictors should not be included unless they were specified in the a priori analysis.

Achen (2005:338) also advocates careful data analysis, but suggests that adhering to 'strict mechanical rules and procedures', (such as to only include three predictors, or conversely, include all attributes as predictors) is not useful. Achen suggests splitting datasets into statistically meaningful subsets (either by theory or analysis), thereby removing the need for many dummy and control variables, and producing homogeneous groups upon which small, coherent regression analyses can be performed.

Taking a different perspective to much of the literature, Berk et al. (2017:2) suggest that the solution is to acknowledge specification problems (i.e. that the model is wrong) where they exist and, rather than abandon regression, make the most of misspecified models; that researchers 'can recognise and accept that requisite assumptions are not met and that the empirical results derive from a misspecified model'. The authors suggest that the 'wrong model' perspective be adopted; that is all models are wrong, but some 'will be more instructive, complete or interesting than others' (Berk et al., 2017:21). The model can be used to describe the dataset at hand, but not in order to make wider causal inferences. Since standard errors, and confidence intervals, etc. are then acknowledged to be wrong, methods such as the bootstrap and 'sandwich' estimator can be utilised to calculate new standard errors and confidence intervals, etc. which are more reliable.

In summary, whilst the focus of much of the critical literature appears to be aimed solely at social science research, this is most likely because social scientists generally work with very complex, inter-related data that may not always be very linear and it can therefore be particularly difficult to satisfy the regression assumptions in these circumstances. Where regression assumptions can be adequately satisfied, there is no reason why linear regression should not be utilised. As has been pointed out in some of the literature, what is missing in some cases is an acknowledgement that regression assumptions (interactions, homogeneity, etc.) have even been considered. And even where assumptions are not satisfied a model may still be useful for describing the particular dataset, as a form of exploratory data analysis. It is likely that many researchers do realise that where regression assumptions are not satisfied their model may not be reliable (particularly for producing wider causal inferences), but for various reasons (unfamiliarity, unsuitability, etc.) do not use alternative methods. The overall consensus is that, where the misuse of linear regression is concerned, deeper data analysis, greater awareness of assumptions, and consideration of alternative techniques may aid in producing more reliable research.

2.7 CONCLUSION

This chapter has explored the background and provided an explanation of OLS linear regression. It has highlighted that, where regression assumptions are satisfied, linear regression can be a powerful tool for social science research, in that results are relatively easy to understand and relationships within the data may be identified and quantified.

However, the use of linear regression in social science research has received sustained criticism over the years; much of this stems from misuse of the method, and the various misconceptions around its usage. The method itself is not criticised. Many of the problems associated with its usage centre upon the fact that to be effective, linear regression relies upon strict statistical assumptions. However, these assumptions can be so difficult to satisfy that they are frequently not adhered to. This is particularly a problem for social science research, since 'real-life' datasets can be very complex, and it may be difficult to identify problems, such as interactions or heterogeneity within data. Often, these difficulties are simply not acknowledged.

Little attention appears to be paid to the problem of misspecified regression models, yet, if a model is misspecified and initial regression assumptions are not satisfied then wider inferences made from them may not be accurate. This can lead to Type I and Type II errors, resulting in inconsistent research and the production of flawed conclusions. Where regression models are correctly specified there is no reason they should not be utilised. Equally, even when they are misspecified in some way, they can still be a useful tool to explore a particular dataset, without making wider inferences.

To some degree, it would appear that since OLS linear regression is considered one of the standard methods of analysis in some fields, there may sometimes be a reluctance to consider alternative methods. However, much of the literature agreed that a greater concentration on methods such as exploratory data analysis and data visualisation might aid in providing a better understanding of the data (and therefore in identifying any problems). This may aid in determining: whether or not a particular dataset and problem is suitable for regression analysis; whether more robust regression methods might be utilised; or whether a different method altogether might be better employed.

This chapter builds the groundwork for future chapters which consider the use of alternative methods to complement established social science methods.

3 STATISTICAL SIGNIFICANCE AND REPRODUCIBILITY

3.1 INTRODUCTION

Together with regression analysis, Null Hypothesis Significance Testing (NHST) is one of the most commonly utilised, and well-established methods in social science research. It is popular because it is a relatively straight-forward process that enables a researcher to consider whether or not a hypothesis might be true given the available data.

This chapter provides an overview of Null Hypothesis Significance Testing (NHST), together with a description of the critical literature surrounding its use. It covers the positives of its usage, together with a description of the various misconceptions surrounding it, and what rejecting a hypothesis actually means. A discussion on the underlying logic of NHST is provided, as well as consideration of the various suggestions in the literature about what might be done to combat the problems associated with the use of NHST.

An exploration of the literature surrounding the reproducibility of social science research is provided. This considers the factors that contribute to problems with reproducibility, such as misuse of various methods and practices such as p-hacking. Suggestions from the literature of how more reliable research results might be produced are considered, together with an exploration of the various advantages and disadvantages of these methods.

3.2 NULL HYPOTHESIS SIGNIFICANCE TESTING

Hypothesis testing is fundamental to social science, and generally involves using statistical methods on a smaller data sample in order to infer something about a larger population. For instance, one might consider whether a regression coefficient has any effect, or whether two population means differ significantly. Most commonly Null Hypothesis Significance Testing is utilised in an attempt to determine whether an effect is 'statistically significant' or whether it might simply be due to chance.

Null Hypothesis Significance Testing (NHST) involves stating two mutually exclusive hypotheses, the Null and the Alternative:

- H_0: Null hypothesis. For example, $\mu = 0$

- H_A: Alternative hypothesis. For example, $\mu \neq 0$, or $\mu > 0$

Generally, it is the Alternative hypothesis that a researcher is interested in (or may believe to be true) and would like to investigate. This is compared to the Null hypothesis, which the data is tested against. Often, the Null hypothesis is actually a 'Nil hypothesis', that is, the hypothesis of zero effect or no difference (Lambdin, 2012; Wasserstein and Lazar, 2016). For example, the hypothesis that a regression coefficient is zero, or that the difference between two means is zero.

In order to determine whether or not to reject the null hypothesis, the null hypothesis is assumed to be true, and given this, statistical calculations are performed upon the data sample (for example, a t-test). The p-value of the test statistic is used to make the decision. If the p-value is less than a predetermined significance level (α), the null hypothesis is rejected; if it is greater, the null hypothesis is not rejected. Generally, α is set at 0.05. However, this is an arbitrary cut-off point (Nelder, 1999; Gelman and Stern, 2006; Greenland et al., 2016), and any value may be used; lower values such as $p < 0.01$ or $p < 0.001$ are sometimes used. The p-value is the probability of obtaining this or more extreme data, given that the null hypothesis is true. Therefore, if the null hypothesis is rejected (i.e. the p-value is below α), this implies that the results are 'statistically significant' at that level (α), and that they are probably not due to chance alone.

However, obtaining a p-value less than the significance level does not necessarily confirm that the research hypothesis is false, it may simply mean that this particular data is unusual, assuming all test assumptions were correct. The p-value may be small because there was error in the data, or because it was drawn from a non-representative sample. Equally, obtaining a p-value larger than the significance level does not necessarily mean that the research hypothesis is therefore proven true, it simply suggests that this particular data is not unusual if all the assumptions were correct. Furthermore, a large p-value does not necessarily indicate a lack of effect, it can also indicate that the data simply could not discriminate amongst many competing hypotheses (Greenland et al., 2016)

Two types of error can result from a hypothesis test:

- Type I error: Rejecting the Null hypothesis when it is actually true. That is, detecting an effect or relationship that is not actually present (false positive)

- Type II error: Failure to reject the Null hypothesis when it is actually false. That is, failing to detect an effect or relationship that is present (false negative)

The probability of committing a Type I error is the significance level, or α (Lambdin, 2012). Given that this is generally set at 0.05, this means there would be a 5% (or 1 in 20) chance of rejecting the Null hypothesis when it actually should not have been rejected. The probability of committing a Type II error, β, is calculable if certain population parameters are known (n, μ, σ and α). The power of a test is the probability of correctly rejecting the null hypothesis, i.e. it is the probability of avoiding a Type II error, and is denoted by 1 – β (Szucs, 2016). A high-power test is therefore desirable.

3.2.1 Critical Literature Surrounding the Usage of NHST

Despite the prevalent usage of NHST in the social sciences, it has been the subject of much criticism over the years, with the literature going back many decades (for example: Berkson (1938), Rozeboom (1960), Bakan (1966), Lykken (1968), Morrison and Henkel (1970), Cohen (1994), Schmidt and Hunter(1997), Nickerson (2000), Armstrong (2007), Ziliak and McCloskey (2009), and Branch (2014)). All make the point that NHST is often misused and misunderstood, and that it provides little useful information. However, this prolonged criticism appears to have had little effect on researchers; NHST remains one of the most commonly applied methods in social science research (Lambdin, 2012; Perezgonzalez, 2015; Ortega and Navarrete, 2017). And achieving statistically significant results is still generally seen as a pre-requisite for research publication (Lecoutre et al., 2001; Branch, 2014; Vidgen and Yasseri, 2016).

Surveying the literature, one of the main criticisms of the application of NHST is that it is frequently misunderstood. This stems from widespread misunderstanding of what p-values actually mean (Branch, 2014; Greenland et al., 2016). The p-value provides the probability of obtaining this (or more extreme) data given that the null hypothesis is true (Wasserstein and Lazar, 2016), that is, P(Data|H_0). But what is commonly (mis)understood is that a p-value is the probability that the null hypothesis is true given the data (Cohen, 1994; Falk and Greenbaum, 1995; Gigerenzer, 2004; Wasserstein and Lazar, 2016), that is, P(H_0|Data). In almost all cases, this is wrong, P(Data|H_0) does not equal P(H_0|Data). Only in rare circumstances might the two be equal (Falk, 1998). The fact that P(Data|H_0) does not generally equal P(H_0|Data) is demonstrated by taking examples of conditional probabilities and reversing them. To utilise Carver's (1978)

example, the probability that a person was hanged given they are dead does not equal the probability that they are dead given they were hanged, that is P(Hanged|Dead) ≠ P(Dead|Hanged). Carver (1978) suggests that the probability a dead person was hanged is likely to be very low (perhaps 0.1), whereas the probability that being hanged would kill someone is probably very high (perhaps 0.97) – to think the two could be equivalent makes no sense. Engman (2013) also provides a hypothetical example utilising Bayes Theorem to illustrate that the probability of concluding that a child has reading inadequacy given they have reading adequacy (0.05) is not the same as the probability of a child having reading adequacy given a conclusion of reading inadequacy (0.22). This mistaken belief is referred to as the inverse probability error (Cohen, 1994; Engman, 2013; Ortega and Navarrete, 2017); and Engman (2013) suggests it is prevalent in sociology research. Whilst basic, both examples highlight that confusing the two probabilities is likely to be misleading.

Another focus of the critical literature is the arbitrary nature of the significance level, α. It would seem that the most commonly utilised value, 0.05, is sometimes viewed as some magical level, yet there is no scientific method behind this choice. It was chosen simply for convenience in the early 1900s when calculations were made by hand using statistical tables and therefore only a limited set of values were available (Boslaugh, 2013:65). Now that calculations of this type are no longer necessary, one could choose any level. The use of a fixed level, such as 0.05, 'promotes the seemingly nonsensical distinction between a significant finding if p = 0.049 and a nonsignificant finding if p = 0.051' (Johnson, 1999:765). Gelman and Stern (2006) state that even a minor (statistically insignificant) change in data can have a large effect upon the p-value produced (despite there being no change in the underlying relationships in the data), and this means an insignificant change can have a profound effect upon whether a null hypothesis might be rejected or not. This highlights that the use of any fixed significance level to make a decision can be damaging since 'changes in statistical significance are often not themselves statistically significant' (Gelman and Stern, 2006:328).

Yet another focus of the critical literature is that NHST is sometimes misapplied. NHST assumes that the data is taken from a random representative data sample, yet many studies do not meet this basic criterion (or fail to correct for it) and this renders any inference about the larger population meaningless (Cohen, 1994; Leahey, 2005). There is

also a tendency to perform NHST on data that equals the population (i.e. data which is not a sample), which also renders NHST meaningless since in this case there is no larger population to infer to, and the sample statistics are therefore equivalent to the population statistics (Leahey, 2005; Engman, 2013). Another important consideration is sample size; like many other statistical measures, NHST is sensitive to the size of the data sample. A small sample size may lead to important effects going undetected, whereas a large sample size may lead to even trivial effects producing very low p-values (Levine et al., 2008)

A further problem raised by the critical literature is that NHSTs are often rendered pointless because the Null hypothesis is known in advance to be false (Bakan, 1966), and given a big enough sample size virtually all null hypotheses will be rejected anyway (Rozeboom, 1960; Thompson, 1993). In the case of the Nil Hypothesis (zero effect) there seems little point in even testing as the null will always be false (Berkson, 1938; Cohen, 1994). As Tukey (1991:100) stated, it is always possible to find some difference in effect:

> All we know about the world teaches us that the effects of A and B are always different – in some decimal place – for any A and B. Thus asking "Are the effects different?" is foolish.

Therefore, there seems little to gain and nothing new to be learnt from rejecting a hypothesis which is known in advance to be false (Berkson, 1938; Armstrong, 2007). Another consideration is that, if the Null is known to be false, then the Type I error rate is actually zero, as it would be impossible to make a Type I error in this case. This would mean that the Type II error rate (one minus the statistical power) becomes the overall error rate. Given that various analyses of social science research publications have consistently found that they lack adequate statistical power (Sedlmeier and Gigerenzer, 1989), this means that the error rate would be much higher than should be acceptable. Schmidt and Hunter (1997) suggest that as a rough average it would be about 50%, and in this case, such a low level of accuracy could be achieved simply by flipping a coin.

These and other misunderstandings of what a p-value actually is have led to various misconceptions about NHST; they are covered extensively in Carver (1978), Cohen (1994), Schmidt and Hunter (1997), Nickerson (2000), Branch (2014), and Greenland et al. (2016). They include believing that:

- the p-value is the probability of the Null hypothesis being true

- the p-value is the probability the results were due to chance
- 1-p is the probability the alternative hypothesis is true
- 1-p is the probability of replication
- a small p-value indicates the results are replicable
- the p-value indicates the importance or size of an effect
- if an effect or relationship is not found to be 'statistically significant' then it is instead zero
- the belief that statistical significance (or rejecting the null hypothesis) means practical or theoretical significance
- the belief that not rejecting the null hypothesis is equivalent to demonstrating it to be true

None of these beliefs are true, however taking this long list of misconceptions into consideration, it seems that the main problem with NHST is that it is sometimes viewed as a test that can answer so much, but in reality, tells a researcher very little.

3.2.2 The Logic of NHST

Berkson (1938), Cohen (1994), Falk and Greenbaum (1995), Schmidt and Hunter (1997), Hofmann (2002), Orlitzky (2012), and Szucs and Ioannidis (2017) argue that it is not just misconceptions that make NHST unreliable, but that there is a far greater problem in that the underlying logic of it is flawed. Superficially, it appears to be based on the *Modus Tollens* logical form, which is denying the antecedent by denying the consequent:

P1: If p (the null hypothesis is true), then q (these data cannot occur)

P2: Not q (these data have occurred)

C: Not p (Therefore, the null hypothesis is false)

This logic is formally valid if the conclusion (C) must be true whenever it's premises (P1 and P2) are true (Hofmann, 2002). However, the problem with NHST is that the reasoning is probabilistic rather than absolute, and by making it probabilistic it becomes invalid (Cohen, 1994; Hofmann, 2002). To quote the example by Cohen (1994:998), in the absolute form, the following is valid logic (if one believes Martians exist):

P1: If a person is a Martian, then he is not a member of Congress

P2: The person is a member of Congress

C: Therefore, he is not a Martian

However, to continue the Cohen (1994:998) example, if one of the premises (P1) is not true, this still leads to a formally correct *Modus Tollens,* but it is no longer logically sound:

P1: If a person is an American, then he is not a member of Congress (WRONG!)

P2: The person is a member of Congress

C: therefore, he is not an American

This example can be made sensible by making it probabilistic, but in doing so it then becomes formally incorrect and leads to a conclusion that is not sensible (Cohen, 1994:998):

P1: If a person is an American, then he is probably not a member of Congress (TRUE!)

P2: The person is a member of Congress

C: therefore, he is probably not an American

Making the *Modus Tollens* probabilistic allows for the possibility of C being false even if P1 and P2 are true, which therefore violates formal deductive logic, as this posits that C must be true when P1 and P2 are true (Orlitzky, 2012). Hofmann (2002:70) suggests that the example shows that NHST is based on 'a faulty conceptualisation of logic', and whilst it might sometimes lead to sensible conclusions, it may also result in wrong conclusions. However, Cortina and Dunlap (1997:166) concluded that 'the typical approach to hypothesis testing does not violate the relevant rule of syllogistic reasoning to any great degree', and quoted different examples that did not cause the *Modus Tollens* to break down. They suggested that whilst Cohen's (1994) example was useful (in highlighting that the application of *Modus Tollens* to probabilistic statements can cause problems), that where sensible research questions were asked (such as the type used in psychology research) then the logic can hold and be useful. Hagen (1997:22) also countered the criticism by stating that arguments can be reasonable even when they are not logically valid, suggesting that in real life most of the decisions we make are 'based on probabilistic premises, not on logic that is valid in a formal sense'.

3.2.3 Suggestions from the Literature on How to Deal with Some of the Problems Associated with the Use of NHST

Given the various misconceptions about NHST, it would seem that one solution might simply be to provide better training in how to use and interpret them correctly. However, whilst it is likely that improved knowledge would be beneficial, it is not always the case that NHST are misapplied or misunderstood; NHST is often implemented correctly. The more fundamental point is that, even when implemented correctly, on their own they do not reveal useful information (Schmidt and Hunter, 1997; Armstrong, 2007; Ziliak and McCloskey, 2009). They should be accompanied by other statistics as in isolation they provide no information about the size of an effect, or the degree of uncertainty around a decision.

Perhaps something that makes NHST so appealing to researchers despite its proven limitations and the steady stream of criticism over the years is its black and white nature – it appears to provide a clear, automatic decision on whether something is 'statistically significant' or not. It provides the perception of a complicated mathematical procedure that results in a definitive answer (Carver, 1978). And in doing so, it makes a researcher's job much easier; simply click a few buttons, check a p-value and a decision is made (Lambdin, 2012). To quote Bakan (1966:430) NHST has:

> removed the burden of responsibility, the chance of being wrong, the necessity for making inductive inferences, from the shoulders of the investigator and placed them on tests of significance

Yet it is precisely this black and white certainty that is dangerous, the NHST is not designed to provide such a definitive answer (Tukey, 1991). NHST should not be used to corroborate theories or hypotheses, to decide on publication, or to make conclusions, it should be the least important part of an analysis (Lykken, 1968). Gigerenzer (2004) suggests that NHST impedes researcher's intelligence and prevents proper statistical thinking. Similarly (Gross, 2015) suggests that its use can result in distracting a researcher from what they are actually measuring and can encourage weak hypothesis testing, and a fixation upon p-values. Kirk (2003) suggests that rather than focussing on p-values the real goal should be deciding whether the data support the scientific hypothesis, the magnitude of any effect and whether it is practically significant. Another side effect of the concentration on 'significance' is that research may be disregarded just because a p-

value is not below some arbitrary threshold; yet this does not mean that an effect is not relevant or may not be of interest to other researchers (Nelder, 1999).

Ward et al. (2010) and Hill and Jones (2014) suggest that if more attention was paid to evaluating model fit, rather than finding 'significant' relationships this may improve theoretical explanations and policy decisions. In particular, analysis of whether attributes identified as significant actually improve the overall fit (or predictive power) of a model could be informative, as insights from a model that fits the data better should be more useful. Lo et al. (2015) state that regardless of the data or problem type, attributes that are identified as significant do not automatically make good predictors. This point is illustrated by Hill and Jones (2014) who found that few of the predictors identified by the literature as important causes of repression were able to improve the predictive power of statistical models of repression; and their work, which utilised cross-validation and decision trees, identified other factors which had so far received little attention in the literature but which did improve predictive power (and therefore warranted further research).

Ward et al. (2010:363) conducted a 'side-by-side' comparison of the statistical significance and predictive power of the different attributes used in two of the 'most influential' models of civil war. They found that whilst the inclusion of some attributes that were considered statistically significant improved the predictive power of the models, others had very little impact, and some actually reduced the ability to make correct predictions. Welch and Goyal (2007:1505) evaluated each attribute using the same methods (generally, linear regression models) and found that attributes identified as being significant (by the academic literature over thirty years) in terms of stock market fluctuations had little predictive power and that 'the profession has yet to find some variable that has meaningful and robust empirical equity premium forecasting power'. It seems counterintuitive to think that attributes which are considered statistically significant may not necessarily provide meaningful improvement in a model's predictive power; however, reasons for this might include model misspecification (such as unidentified interactions) and failure to test models on out of sample data (Ward et al., 2010). However, the examples highlight that the consideration of statistical significance alone can be misleading and may not necessarily lead to developing models with useful predictive power (Ward et al., 2010; Hill and Jones, 2014).

Having 'significant' results is seen as necessary for research publication in many fields, and reliance upon significance testing is so strongly embedded in researcher's minds and habits that alternatives are met with strong resistance (Schmidt and Hunter, 1997). Many researchers use NHST simply because that is the method they have always used. Cohen (1994), Falk and Greenbaum (1995) and Gigerenzer (2004) all refer to the 'illusion' of NHST, and suggest that the reason it is so deeply embedded in researcher's minds is that there is a ritualistic nature to it that is perpetuated by social pressure and wishful thinking. Researchers may know that some of their beliefs about NHST are not true, but they prefer to act as if they were true anyway; to quote Cohen (1994:997) 'it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!'

Ziliak and McCloskey (2009:2302) are particularly harsh in their criticism, suggesting that reducing scientific problems to an interpretation of 'statistical significance' is not scientific and has had a detrimental effect upon society as a whole:

> Statistical significance is, we argue, a diversion from the proper objects of scientific study. Significance, reduced to its narrow and statistical meaning only—as in 'low' observed 'standard error' or '$p < .05$'—has little to do with a defensible notion of scientific inference, error analysis, or rational decision making. And yet in daily use it produces unchecked a large net loss for science and society. Its arbitrary, mechanical illogic, though currently sanctioned by science and its bureaucracies of reproduction, is causing a loss of jobs, justice, profit, and even life.

They, along with Carver (1978), Cohen (1994), Schmidt and Hunter (1997), Gigerenzer (2004), Armstrong (2007) and Cumming (2014) suggest that researchers stop using NHST altogether, arguing that it is pointless. Much of the other critical literature, whilst stopping short of suggesting a ban, argue that the use of a decisive accept or reject statement is wrong and discourage the use of NHST as the only method of evaluation. They suggest NHST might still be considered, albeit with a far less prominent role, alongside other statistical methods and that it should be accompanied by statistics such as confidence intervals and effect sizes (Kirk, 2003; Levine et al., 2008; Gross, 2015). These statistics might help convey information about the magnitude of an effect and whether it is relevant. Szucs and Ioannidis (2017) suggest that NHST should no longer be the 'cornerstone', or automatic default method of research; it's usage should be clearly justified (with alternative methods considered), and that researchers should pre-register

hypotheses and analysis parameters in order to focus more completely on the particular research question. Van de Schoot et al. (2011) suggest using NHST in a more intelligent way and considering informative hypotheses (rather than nil hypotheses) that might actually be true. The general consensus in the literature is that more intelligent data analysis is required. And much more importance must be placed upon replicability of results (Falk, 1998; Schmidt, 2009; Cumming, 2014). Lambdin (2012) and Branch (2014) suggest that once other statistical methods are used more frequently, NHST might eventually be seen as unnecessary and phased out.

It is notable that the amount of literature critical of NHST far outweighs any positive literature; there seem to be few research articles actively supporting the usage of NHST. However, as already stated, there is substantive literature that overwhelmingly uses NHST and would therefore seem to tacitly support its use. There is a small core of literature that defends NHST from the methodological criticism, however, even they generally acknowledge the various limitations and do not recommend it be used in isolation (Abelson, 1997b; Hagen, 1997; Chow, 1998; Levin, 1998; Nickerson, 2000; Mogie, 2004). Most argue the issues are around misconceptions and misuse by researchers, rather than flaws in the process. Hagen (1997) asserts that critics have used extreme examples to argue their case and that NHST is 'unfairly maligned', whereas Abelson (1997b) suggests that misunderstandings happen with many statistical measures and are not unique to NHST.

Abelson (1997a:117) believes the fact that NHST provides a categorical statement (accept/reject) actually stimulates further research, and that 'Significance tests fill an important need in answering some key research questions, and if they did not exist they would have to be invented'. Levin (1998) suggests that NHST might be used in a more intelligent way, with carefully developed hypotheses, control of Type I errors and effects sizes, and optimal sample sizes. Whilst Mogie (2004) believes that NHST can provide a clear answer to well formulated questions, but that it should be complemented by other statistics. Chow (1998) suggests that NHST should not be used to corroborate theory, but instead to exclude chance, and acknowledges that NHST says nothing about real-life importance.

One reason for the overall lack of supportive literature may be that NHST is such an embedded technique that researchers feel it does not need to be defended – if it is

widely, and actively used, and in textbooks, then there is no need to defend it. Another reason may be that there simply is no way to defend many of the criticisms. Another possible reason may be what Nickerson (2000) touched briefly upon; the idea that the so called 'significance controversy' may be something that exists mostly within the group of authors who write about it, and that those outside this group may be unaware of the limitations. Following this line of thought, the various statistical textbooks have over the years done little to warn students and researchers of the weaknesses and misconceptions surrounding NHST (Gigerenzer, 2004).

Perhaps one reason why the usage of NHST is still commonplace despite the criticism is that it is often falsely perceived as the only objective approach to scientific inference and 'alternatives are simply not taught and/or understood' (Szucs and Ioannidis, 2017:14). As Wasserstein and Lazar (2016) point out, there is a circular logic to the continued use of NHST; it is taught because that is what the scientific community and journals use, and it is still used because that is what people were taught. It is therefore difficult to break the cycle, and this difficulty is compounded by the fact that there is no obvious replacement – there is no test that can provide an automatic definitive answer. As Cohen (1994:1001) stated: 'First, don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist.' Given this, at least some of the literature (both critical and positive) agree that there might still be a place for NHST, but that hypotheses should be very clearly stated, exact p-values should be reported and not used to provide a decision, and that they should be accompanied by other statistics such as effect sizes and confidence intervals.

3.3 REPRODUCIBILITY

As has been discussed in previous sections, research practices such as failure to satisfy regression assumptions, improper evaluation of models and the misuse of methods such as NHST can lead to inconsistent research results. This ultimately undermines overall research quality, as a lack of reproducibility can lead to problems such as blindly accepting results that are wrong, doubting results (that may be correct), and in general, the production of conflicting conclusions (Ioannidis, 2005; Cumming, 2014). Overall, this 'wastes research funding, erodes credibility and slows down scientific progress' (Szucs, 2016:1). Ioannidis (2005) claimed that most published research findings are actually false; reasons for this include the use of NHST, bias, studies with low power, data

dredging and selective reporting of results. Whilst this might seem a dramatic claim, there is no doubt that there are many conclusions drawn from academic studies that later prove to be false (Nosek et al., 2012).

Where multiple hypotheses are tested, satisfying assumptions and interpreting results correctly becomes even more difficult. For instance, when examining large social survey datasets, hundreds or even thousands of attributes might be investigated, with people grouped in many different ways. The likelihood of achieving a statistically significant result increases with every predictor added to a model (Smith et al., 2002), and given twenty predictors, one will be significant at the $p <= 0.5$ level through chance alone (1/20 = 0.05). The probability of at least one Type I error rises rapidly with the addition of more hypotheses (Shaffer, 1995). Benjamini (2010) highlights the problem of multiplicity, and the lack of adjustments for it, leading to false results. There are adjustments, such as the Bonferroni correction, the False Discovery Rate and the Familywise Error Rate that can help. However, a problem with perhaps the most frequently utilised measure, Bonferroni correction (divide α by the number of tests being performed) is that it is too conservative (Perezgonzalez, 2015); and in minimising the risk of Type I errors, the power is reduced, which consequently makes Type II errors more likely (Smith et al., 2002).

3.3.1 P-hacking

Something that is particularly damaging to the reliability of research is P-hacking. This is the process of repeating an experiment until a statistically significant result is obtained. Over the years it has also gone by names such as bias, significance chasing or searching, data snooping, and data fiddling (Simonsohn et al., 2014). Historically it has also been referred to it as 'data mining' (Lovell, 1983). P-hacking can happen (either deliberately or accidentally) when decisions about data are not made in advance. For example, during the analysis process extra data might be collected, or different attributes excluded or included in experiments. As these decisions are being made with prior knowledge of results this may then make the study more likely to have a significant result; the previous experiments are simply filed-away and not mentioned in the final analysis (Simonsohn et al., 2014). The effects of P-hacking mean that the Type I error rate is inflated and studies could reveal significant relationships where there is actually none (Simmons et al., 2011; Szucs, 2016). Simmons et al (2011) suggest that in many cases it is more likely that a

researcher will find false evidence of the existence of an effect, than that they will correctly find evidence that it does not exist.

Simonsohn et al. (2014) suggest creating 'p-curves' in an attempt to identify p-hacking. This involves selecting multiple research studies (perhaps by journal, subject, or a particular finding or hypothesis), then extracting the important p-values and plotting them. Sets of studies with only true effects should generate right-skewed curves (i.e. they contain many low p-values, e.g. 0.01), whereas those where no effect exists should be uniform. Those that have been intensely p-hacked should produce left-skewed curves. The theory is that studies which have been p-hacked are likely to contain p-values that are just below 0.05, i.e. they are just statistically significant, whereas studies with very small p-values are more difficult to obtain and therefore are more likely to contain true effects (Simonsohn et al., 2014).

However, Bruns and Ioannidis (2016) question the reliability of the p-curve when used for observational research (as opposed to randomised studies); stating that biases in the model (misspecification of regression models, measurement errors) can also lead to right-skewed p-curves; therefore, p-curves are unreliable in distinguishing between true effects and null effects with p-hacking. Overall, while p-curves might possibly provide some information about whether there is a particular effect, it is not clear that it would reduce p-hacking or whether it could be trusted. If less focus was placed upon obtaining 'significant' results, it is likely they would not be needed.

Problems with misunderstandings and misuse of p-values seem to pervade social science research; to counter these problems some journals have discouraged their use (Lang et al., 1998). The American Psychological Association considered banning significance tests altogether in 1999, although later produced a set of guidelines instead (Wilkinson and the Taskforce on Statistical Inference, 1999). The journal Psychological Science is now actively screening for studies that may be questionable in terms of replicability, i.e. those that have low statistical power, p-values just less than 0.05 and report surprising results (Lindsay, 2015). The journal of Basic and Applied Social Psychology recently banned the use of NHST (and any inferential statistics) altogether in the hope that requiring strong descriptive statistics would increase the quality of submitted research (Trafimow and Marks, 2015). The discussions surrounding this ban led the American Statistical Association to issue a statement on p-values in which they provided six principles to help

improve understanding of what a p-value is and how it should be interpreted. In brief, they stated that p-values do not measure the probability that a hypothesis is true, that conclusions or decisions should not be based upon whether a p-value passes a specific threshold, and that a p-value alone cannot provide a good measure (Wasserstein and Lazar, 2016). Greenland et al. (2016) in response also highlighted the 'rampant' misinterpretation and abuse of statistical tests, and in particular the harmful usage of p-values to determine 'significance'.

3.3.2 Replication

In other scientific fields, replication is common (and often required) in order to verify research. Replication is seen as the 'cornerstone of science' (Carver, 1978:392; Simons, 2014:76), and is generally required to give credibility to a theory or hypothesis. Yet in many social science research areas findings are rarely replicated and so false conclusions persist (Freedman, 2009; Schmidt, 2009; Nosek et al., 2012; Cumming, 2014). This is likely due to many factors, such as publication bias, selective research, or a lack of time, money or interest. Academics are encouraged to produce novel research, therefore there is little incentive to reproduce research already done (Open Science Collaboration, 2015). And journals publish statistically significant results far more frequently than statistically insignificant (or null) results; there is little interest in negative results (Lehrer et al., 2007; Nosek et al., 2012; Couzin-Frankel, 2013; Franco et al., 2014).

Negative results are becoming less frequent in published research. Fanelli (2012) found that between 1990 and 2007 there was an overall increase of over 22% in the proportion of papers reporting a statistically significant effect; this was even higher for the social sciences. Franco et al. (2014) suggest that significant results are 40% more likely to be published than null results in social science research, and because of this authors often do not even write up findings that were null. The selective publication of results is often referred to as the 'file drawer problem'. Research seen as unproductive is simply filed away (Rosenthal, 1979). The problem with this system is that published results may reflect an increased likelihood of Type I error (false positive), and hidden null results mean that the research community as a whole is excluded from expanding its knowledge of a research area (Fanelli, 2012; Franco et al., 2014).

An example of the overall lack of reproducibility in social science research was highlighted by the Open Science Collaboration (2015) which replicated 100 experimental and

correlational studies taken from three top Psychology journals. Original data was used where possible and the authors were contacted to clarify methods. 97% of the original studies had significant results ($p < 0.05$), whereas only 35% of the replicated studies did. However, none of the original studies were contradicted, but the replicated results were statistically weaker. With such a project, it is important to consider that even where the ideal replication is performed, even an ideal study may fail to replicate sometimes; and conversely, that a failure to replicate does not necessarily indicate flaws in the original study (Lindsay, 2015). However, even accounting for this, the Open Science Collaboration results highlight not only the difficulty of replicating research, but the fact that many studies produce results that simply do not stand up to further scrutiny.

It is possible that errors in research might be reduced by moving towards a framework of reproducible research, i.e. by requiring researchers to provide the data, methods and code used to produce results (Fomel and Claerbout, 2009); at the very least, a more open research community would likely increase discussion and improve overall standards (Nosek, 2015). Researchers should be encouraged to find flaws in their work, and Alberts et al. (2015) suggest that in order to improve the quality of academic work there should be incentives for publishing good quality work rather than for the amount of work produced (citations, etc.).

It is also suggested (Schmidt, 2009; Nosek et al., 2012; Cumming, 2014; Franco et al., 2014; Vidgen and Yasseri, 2016; Szucs and Ioannidis, 2017) that pre-registration of studies may help to improve overall research quality. Information such as the research questions and objectives, methods to be utilised and sample sizes might be registered in advance; deviation from these would need to be justified. (Nosek et al., 2012:625) make the point that a registry could help distinguish between chance discoveries (which may be less reliable) and prior predictions which were confirmed by the study (and likely to be more reliable), stating that 'the point of making a registry available is not to have a priori hypotheses for all projects and findings; it is to clarify when there was one and when there was not. When it is a discovery, acknowledge it as a discovery'. Another advantage of pre-registration could be that a researcher might know in advance where a particular hypothesis has failed previously and therefore may consider whether to spend their time researching it; previously such null results may have been hidden. However, a concern with pre-registration would be that it could stifle more exploratory work, Gelman and

Loken (2014:464) suggest that 'the most valuable statistical analyses often arise only after an iterative process involving the data' and that pre-registration may be practical in some fields and for some problems, but that it cannot be a general solution. Another concern with this method could be that researchers may not want to disclose their hypotheses (if for instance, they feel their idea may be copied), however, both Cumming (2014) and Szucs and Ioannidis (2017) makes the point that pre-registration need not be public.

Sharing data and code is also suggested, and in theory, the idea of requiring researchers to provide their data, and code, etc. appears very useful, but it might not always be feasible: data may not be shareable for ethical or legal reasons; it may be impractical (or expensive) to store, or in an unusual format; the cost of producing accompanying documentation could be prohibitive; researchers or institutions who have put great effort (and money) into building a dataset may not be inclined to share it; and in a competitive research community, researchers may not want to allow access to their data in case someone else beats them to a useful discovery. Abbott (2007) also makes the point that making data more freely available may lead to less respondents to surveys etc., since they may be put off by the thought of their information being more freely available. Given any of these reasons, it therefore may not always be realistic to share data; however, Freese (2007) suggests that at the time of publication researchers should state everything that would make replication easier, and if they cannot provide data this should be transparently explained.

Another method that would improve reliability of results, and which is underused in the social sciences, is the utilisation of cross-validation when building models (Freedman, 2009; Hindman, 2015; Woodside, 2016). Since most social science researchers use all of their data to fit a model, it is then difficult to tell whether their results are useful or simply a peculiarity of the particular dataset (Hill and Jones, 2014; Hindman, 2015). Testing how a model will perform on previously unseen data can provide insight into the generalisability of a model and is common practice in other scientific fields. If a model has captured the underlying relationships within a dataset, then it should perform well on new data; if it has simply captured relationships within this particular dataset (overfitting) it will not perform well on new data (Ward et al., 2010; Hill and Jones, 2014). It would be unthinkable in the field of machine learning, for example, for a model not to have some form of cross-validation applied to it. Greater utilisation of cross-validation could be a

simple way of automatically helping to validate models, and therefore providing more research credibility. Freedman (2009:75) suggests that whilst cross-validation is not as good as real replication, it is still much better than nothing. One argument is that cross-validation may be impractical where extremely small datasets are concerned, however, methods such as Leave-One-Out-Cross-Validation, which require only one record to be excluded as a test case, or bootstrapping, which samples with replacement, are methods that would not 'waste' potentially expensive data. Cross-validation is considered further in section 4.5

Overall, it seems that there can be a tendency in social science research to use the more established methods (regression, NHST), even where their use may not be suitable; or conversely, it may be that their use is suitable, but often the expectations around their usage are misguided. The literature has highlighted that when regression models are misspecified, or where too much emphasis is placed upon obtaining significant results, it can lead to unreliable, or inconsistent research. Taagepera (2008) suggests that this limitation in social science methods, means that much research is often never used again once published and therefore may have little impact upon the real world.

Much of the critical literature suggests that, where feasible, moving towards a system where social scientists are asked to pre-register, or provide the basic tools for replication, might produce more reliable results and improve overall standards. However, this may not always be practical, therefore the focus should also be on ensuring that less importance is placed upon significance testing, and that other methods and measures are also considered.

3.4 CONCLUSION

Along with linear regression, NHST is one of the most commonly utilised methods in social science research. This chapter explored the usage of NHST in the social sciences, and considered the literature surrounding the various criticisms and the problems associated with its use. It highlighted that NHST is often misunderstood and misapplied, and that placing too much importance upon the results of NHST can distract from providing statistically sound analysis. The dichotomous nature of the NHST, so often reduced to an accept or reject decision, and based upon an arbitrary significance level can encourage complacency in research.

In some fields, achieving a statistically significant result is still generally considered a pre-requisite for research publication, and so the use of these tests may still be necessary. Whilst there are suggested alternatives to NHST, such as providing other descriptive statistics, the literature highlighted that there is no equivalent alternative test. Or at least, no alternative that fulfils the incorrect perception of NHST. That is, there is no reliable replacement test that can provide a definitive yes or no answer about a hypothesis.

Overall, this and the previous chapter have highlighted that there are some disadvantages to the current, more established, social science methods. Where methods are misused, regression assumptions are not satisfied, or too much emphasis is placed upon p-values and obtaining significant results, this can lead to research that is unreliable. This can mean that research results as a whole are less trusted and undermines research quality. This lack of reproducibility can mean that research results may be wrongly accepted, or (perhaps wrongly) doubted, or simply lead to conflicting results. Suggestions in the literature to combat these problems include the pre-registration of studies and making efforts to provide the data and code used. More reliable research might also be achieved by considering alternative methods more commonly utilised in data mining, and this is considered in the following chapters.

4 DATA MINING

> Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise (Tukey, 1962)

4.1 INTRODUCTION

Data Mining is the process of discovering hidden patterns or previously unknown relationships in data sets – it can be used to cluster, classify, derive rules and associations, make predictions, detect anomalies, and summarise and visualise data. It combines traditional statistical and database techniques with methods from the fields of machine learning and artificial intelligence to produce (sometimes) sophisticated algorithms and methods. Data mining is often referred to as Business or Predictive Analytics in the business world, and more recently the terms Machine Learning, Data Science and Big Data have been increasingly used to describe data mining – but whatever the name, or goal, all methods are essentially trying to make some sense out of data.

In the social sciences, the term Data Mining has generally had negative connotations over the years. Historically it was viewed as the process of repeating experiments and/or fiddling data until a statistically significant effect was found (also known as p-hacking, data dredging, etc.). However, it does not have this meaning outside the field of social science, in other fields it is synonymous with machine learning, and data science, etc. In this thesis, the term Data Mining is used to refer to the whole process involved in a project (data exploration, cleansing, visualisation, etc.), whereas Machine Learning refers specifically to the methods used (such as decision tree learning, or clustering). That is, whilst the two terms may sometimes be used interchangeably, in this case machine learning is considered one part of the overall data mining process.

This chapter gives a brief explanation of data mining, how it developed and why there is a need for it. An explanation of the general processes is provided; of supervised and unsupervised learning, model evaluation, and the use of cross-validation to more rigorously evaluate models. There is a description of clustering methods and decision tree learning, random forests and boosted models, as these are methods that are utilised in the Case Study chapters. They are also considered because they are methods that are more interpretable and so perhaps more useful to social scientists. As with regression

methods, there are negative aspects associated with the use of data mining and these are considered. This chapter provides the background for the following chapter which considers the use of data mining in social science research.

4.2 BACKGROUND

Humans have long attempted to identify patterns and make sense from data: organising objects into logical groupings is fundamental to human understanding (e.g. biological classification); and looking for relationships within data and ways to gain insight and predict future behaviour is nothing new (e.g. linear regression). Data mining has origins in statistics and mathematics, and many of its methods were derived long before the term 'data mining' was coined – for example, linear regression (early 1800s), k-means clustering (1960s), and artificial neural network methods (1940s). In essence, the term data mining is simply used to draw together many techniques, both old and new.

In 1962 John Tukey (1915-2000) considered the future of data analysis, stating that although there had been great advances in statistics over the last century this had not had a corresponding effect upon data analysis. He felt that data analysis should be considered a science in itself, and that to make advances it would be necessary to tackle more realistic problems and move away from using rigid statistical assumptions. He believed that it was not always possible to derive an exact solution to a problem, as real-world data often does not fall into neat mathematical distributions. Rather real-world data might require a solution to be more approximate, but that this was surely better than an exact result based upon false statistical assumptions (Tukey, 1962). Much like the data mining methods used today, Tukey felt that data analysis should be a flexible and iterative process, and that it made sense to study what worked and what did not within data analysis and then adapt accordingly. He saw the need to improve methods for the treatment of incomplete and spotty data (i.e. data containing outliers, errors, and non-Normal distributions) and that better graph plotting techniques were required (Tukey subsequently invented the box-plot (Tukey, 1977)).

In 1977 Tukey published Exploratory Data Analysis (EDA), a guide to exploring data, with particular emphasis placed upon graphical methods. He felt that statistics placed too much importance upon hypothesis testing, which he called confirmatory data analysis (CDA), and that EDA should always be the first step in model building. He believed that

greater emphasis should be placed upon using EDA to suggest hypotheses to test, and that it could also be used to assess which statistical techniques might be appropriate for the data, and to decide whether further data should be collected (Tukey, 1977). Although written before the age of widespread computer usage, many of Tukey's EDA techniques are still used by data scientists today.

Aside from its initial pejorative usage (Lovell, 1983), the term 'data mining' began to be used within the academic research community in the late 1980s (Coenen, 2011). Around this time, relational databases had come into use, meaning data could be stored in greater amounts, and was more easily accessible, than ever before. In 1989 the first Knowledge Discovery in Databases (KDD) workshop was held, and in 1995 it became an annual conference (the ACM SIGKDD Conference on Knowledge Discovery in Databases and Data Mining). In 1997 the Data Mining and Knowledge Discovery journal was launched, highlighting the increasing usage of data mining. Many more journals followed suit, and over the last 20 years, articles related to data mining have been published in over 2000 different academic journals (Web of Science, 2017).

Concurrently, in the business world in the 1980s, the field of Database Marketing, which is utilised to analyse customer databases and examine customer preference, had also begun to employ data mining methods. Large companies (such as GM, AT&T and Kraft) were beginning to gather data and utilise their massive databases to make predictions at an individual-level, rather than, as previously, for aggregated groups of people (or market segments) (*Database Marketing - Businessweek*, 1994). This allowed them to consider how likely customers were to buy their products and then market items specifically at the individual. Over the intervening years, these methods have become commonplace and businesses routinely use data mining techniques to look for a competitive edge over their rivals.

Over the last two decades, our ability to store data has risen at an exponential rate, driven by technological improvements and coupled with a reduction in the cost of data storage. In 2012 it was estimated that 2.5 exabytes (2.5 billion GB) of data are created each day, and that this number will double every 40 months or so (McAfee and Brynjolfsson, 2012). Between 2013 and 2020 the 'digital universe', which is all the data created in a single year, is predicted to grow by a factor of 10 - from 4.4 zettabytes to 44 zettabytes (44 trillion GB); and every two years it more than doubles in size (IDC, 2014).

Data is being stored from every imaginable source - for example, customer databases, social media, sensor data, financial data, governmental data, and biomedical and scientific data. Many traditional statistical techniques are simply not equipped to cope with the size, type and dimensionality of these large quantities of data. The development of data mining was driven by the need for new techniques, and fuelled by ever increasing computational power. It would be impossible for humans to manually analyse this increasing volume of complex data, much of it would remain untouched (much of it still does). It is likely that without the development of data mining techniques, hidden patterns in data and valuable insights would remain undiscovered in these large datasets.

4.3 THE DATA MINING PROCESS

There have been attempts to formalise the process of data mining, perhaps the two most notable are Knowledge Discovery in Databases (KDD) and The Cross Industry Standard Process for Data Mining (CRISP-DM). KDD, in Figure 2, was developed from an academic viewpoint and the steps in the process are: understanding the problem; acquiring and selecting data; cleaning, pre-processing and transforming the data; data mining; and finally, interpretation and evaluation of the results (Fayyad et al., 1996).

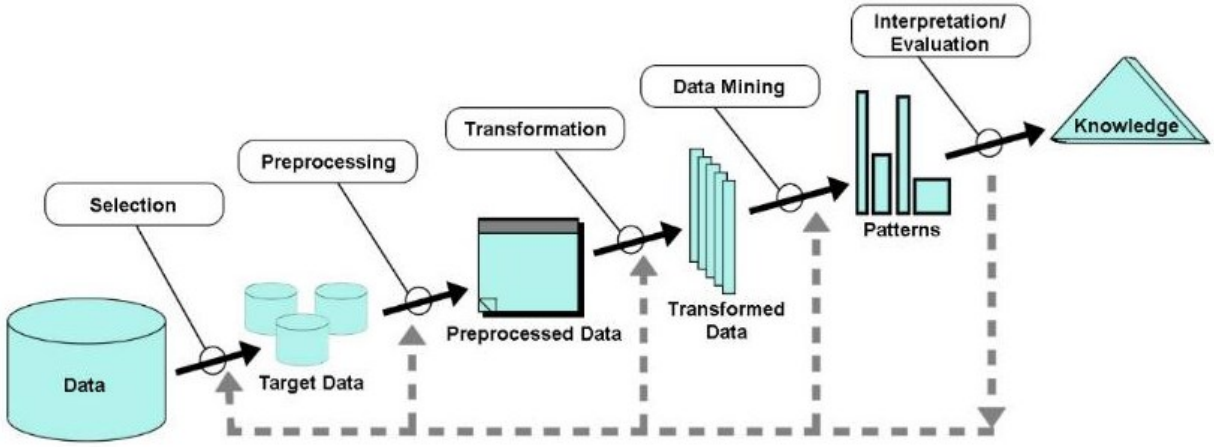


Figure 2: The Knowledge Discovery in Databases (KDD) Process as defined by Fayyad et al. (1996)

CRISP-DM, in Figure 3, was developed by three large companies (DaimlerChrysler, SPSS and NCR) and presents a more business oriented approach, adding a Business Understanding element to the process (Chapman et al., 2000). Two others methodologies exist: SEMMA (Sample, Explore, Modify, Model and Assess) which was developed by the statistical software company SAS for their software (Azevedo and

Santos, 2008); and more recently the Big Data Analysis Pipeline which was published in the Computing Community Consortium Big Data Whitepaper (Jagadish et al., 2012).

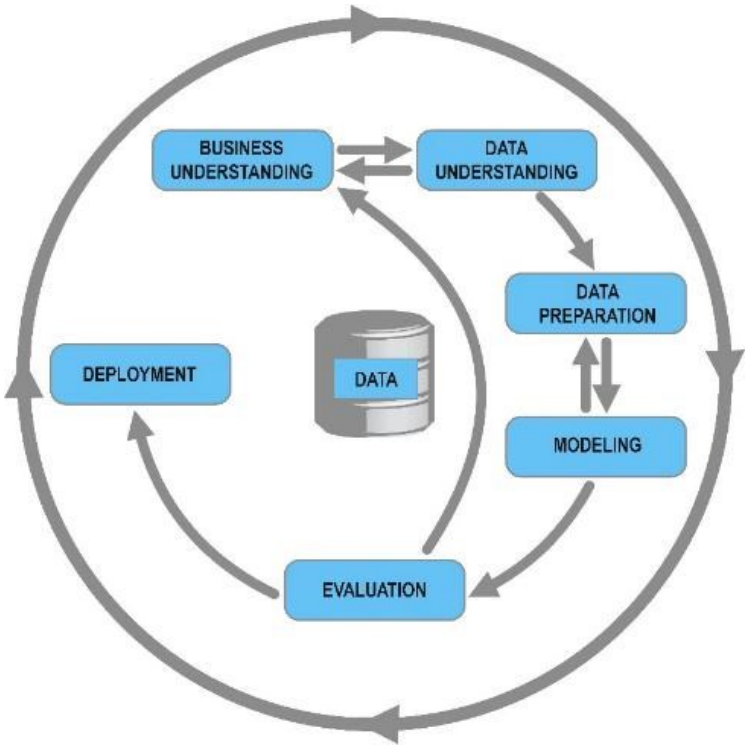


Figure 3: The CRISP-DM Process, as defined by Chapman et al. (2000)

Despite their development by entirely different groups of people, years apart and with differing goals (business, academic, etc.), these processes are all strikingly similar. They are iterative and the data mining, or modelling step, is simply one part of a larger process. The different methodologies all highlight that data mining (or the modelling, or machine learning step) cannot be performed in isolation; even in the simplest situation there would usually be a need for an understanding of the problem, and some form of data preparation and exploration before deciding upon a suitable algorithm. The methodologies also highlight that the term 'data mining' is used interchangeably to mean both the whole process and the modelling step. For the purposes of this thesis, the modelling step is generally referred to as 'machine learning' and the whole process is referred to as 'data mining'.

4.4 Supervised and Unsupervised Learning

Broadly, there are two types of machine learning methods: supervised and unsupervised learning (Witten et al., 2011; James et al., 2013). Supervised methods typically involve predicting or estimating a target attribute based upon certain predictor attributes. They

are ‘supervised’ because the algorithm learns by example, i.e. when training the model, the correct answer is already known. The model may then be deployed on new (unknown) data. Decision trees, artificial neural networks and linear regression are all examples of supervised learning. In the case of unsupervised learning, there are inputs or predictor attributes, but no target attribute, i.e. the data is unlabelled, there is no ‘answer’. Unsupervised learning is used to discover relationships and hidden structure in data, for example in cluster analysis.

Choice of algorithm depends very much on the problem, but can also be dictated by the type of data and the software and hardware that is employed. Some algorithms only work with specific types of data (such as numerical), some are unsuited to mixed data, and some cannot handle missing values. Whilst it is possible to transform data, it may sometimes make sense to choose an algorithm that matches the data. The software or programming environment employed can also heavily dictate which algorithms might be used. Whilst popular statistical programs such as SPSS provide (relatively) simple to use interfaces, they usually contain only a basic selection of data mining algorithms and can often take many years to integrate newer algorithms into their software (if at all). Open source environments such as R or Python generally contain a wider choice of algorithms; new and experimental algorithms are freely available since anyone may create them. However, there may be a steeper learning curve associated with their use, and they tend to have interfaces that are less user-friendly.

4.5 Model Evaluation and Cross-Validation

There are many methods for assessing the validity of a data mining model, for example: classification accuracy; total cost or benefit (different errors might incur different costs); error; lift; ROC curves; or R^2. However, the overriding theme for supervised learning is that the final model should always be tested on new, previously unseen (or out of sample), data. This new data should be data that was not used anywhere else in the model building process. How a model performs on previously unseen data provides an unbiased view of the quality of the model (Breiman, 2001b; James et al., 2013). Any estimate of a model that is judged by the same data it was trained upon is likely to be optimistic - using new data checks that the model will generalise well to other, previously unseen, data (Domingos, 2012).

It is therefore important to perform some sort of model validation when performing supervised learning. Overfitting occurs when a model has learnt the training data so well that it cannot generalise for new data (Breiman, 2001b; Witten et al., 2011; Domingos, 2012; James et al., 2013). If a model is overfitted it may be excessively complex, and rather than identifying patterns in the data, it may simply be describing random error or noise contained in the data. That is, it may have learnt that particular training data set so well, including any unusual features (which may just be a random part of that dataset), that the model does not represent a meaningful, generalizable relationship between the predictors and target attributes (James et al., 2013; Kuhn and Johnson, 2013; Hill and Jones, 2014). This is also often referred to as a model having high variance. Underfitting occurs when a model has been unable to capture the underlying patterns in the data and misses important relationships; this is also referred to as having high bias (Domingos, 2012). Both underfitted and overfitted models do not perform well on new data.

Ideally, a model should have low variance and low bias; that is, it should generalise well on new data and also have accurately captured the underlying patterns in the training data. The problem of minimising both of these factors is often referred to as the 'bias-variance trade-off' (Domingos, 2012; James et al., 2013). Striking a balance can often be difficult to achieve; for instance, it is much easier to achieve a low-variance, high-bias model, or vice versa (James et al., 2013:36).

Perhaps the most important step in building any supervised machine learning model is that the model must be validated on previously unseen, or out of sample, data. In fields where machine learning methods are more readily utilised it would be unthinkable not to validate a model in this way. Yet, in contrast, much social science research does not utilise any model validation on out of sample data (Berk and Bleich, 2013; Hill and Jones, 2014; Woodside, 2016). This means that it is difficult to determine whether patterns or relationships discovered in the data could apply more generally or whether they are simply a feature of the particular dataset (i.e. the model is over-fitted).

Perhaps the simplest way to validate a model is to randomly split the data into a training and testing dataset; the model is built using the training dataset, and then the test data is run through the final model in order to evaluate model performance on new data. An alternative is to use a training, validation and testing dataset; the validation set might be utilised for evaluating differing parameters or settings etc., then the final test of the

model is again performed using the test dataset. A disadvantage of using this 'holdout' method is that it wastes data; if there is a small amount of data then there may not be enough to split into two or three subsets, and the test set might therefore be biased as it is so small (Kuhn and Johnson, 2013).

An alternative is K-fold cross validation which randomly partitions the data into k equally sized subsets. One of the subsets is retained for testing and the model is trained upon the remaining k-1 subsets; this is then repeated k times with each of the k subsets used once as the testing data (Domingos, 2012; James et al., 2013). The k prediction errors are then averaged (or some other metric may be used) to produce an overall estimation of the error (Hastie et al., 2009). Often 10-fold cross-validation is utilised; it appears to be the default setting in many machine learning programmes. However, any number of folds might be utilised; Leave-One-Out Cross-Validation (LOOCV) uses only one record as the test set and the rest of the data for the training set each time, and this is particularly useful for very small datasets. However, a disadvantage of LOOCV is that it has higher variance than k-fold cross-validation (where k is less than the size of the dataset, n) due to each training set consisting of almost the entire dataset; this means that the estimates from each fold can be highly correlated and therefore their average can have high variance (James et al., 2013:183). Overall, the advantages of k-fold cross validation are that it matters less how the data is divided, as all the data is used for both training and testing; and it can be useful when a dataset is small. A disadvantage, particularly when dealing with a large dataset, is that it takes k times as much computational time. Domingos (2012:81) also points out that if it is used 'to make too many parameter choices it can itself start to overfit'.

An alternative to k-fold cross-validation is the bootstrap. Data is selected, with replacement, from the dataset to form the training set. This training set is the same size as the dataset, but some records will be present multiple times, whereas others not at all. The records not selected form the 'out-of-bag' samples. A bootstrap is usually performed multiple times, and for each iteration the model is built on the training set, and the error rate calculated on the out-of-bag samples (Hastie et al., 2009; Kuhn and Johnson, 2013). Bagging, or bootstrap aggregation, is further discussed in section 4.8.1. Where comparing the performance of k-fold cross-validation to the bootstrap, in general, k-fold cross-validation may be prone to large variance (particularly with small datasets), whereas

bootstrapping can reduce the variance but is likely to be more biased (Efron, 1983). However, where large datasets are utilised, issues with variance and bias become less of a problem (Kuhn, 2013:70).

In the case of unsupervised learning, it can be difficult to evaluate model performance as there is generally no 'answer' to compare the results to. The evaluation method may be determined by considering what the goal was, and whether the solution fits that, or perhaps a more general feeling that the model describes whatever it should correctly. Certain techniques do have more concrete evaluation methods; in clustering, for example, there are methods to examine the quality of a clustering (these are covered in section 4.9.3). A clustering might also be judged simply by whether the data falls into what appear to be reasonable groupings, i.e. using prior expert knowledge. Evaluation with previously unseen data can be useful in some situations - if the same method is performed on new data and provides similar results, then this may go some way towards validating the results. However, this uncertainty around the evaluation of unsupervised learning can make its use challenging (James et al., 2013). Hastie et al (2009:487) state that one 'must resort to heuristic arguments not only for motivating the algorithms, as is often the case in supervised learning as well, but also for judgments as to the quality of the results. This uncomfortable situation has led to heavy proliferation of proposed methods, since effectiveness is a matter of opinion and cannot be verified directly.' There is often no 'right' way to evaluate an unsupervised model. However, a particular use for unsupervised methods is in exploratory data analysis – that is, to look for patterns or identify research questions that can then be further explored or validated by other methods.

The importance of validating models cannot be overstated, and these validation methods (k-fold cross-validation, etc.) generally work with any kind of supervised learning model. There is no obvious technical reason why linear regression, which is arguably the most commonly utilised social science method, could not be evaluated in this way more frequently in social science research (Hill and Jones, 2014). It would seem that if validation methods were utilised more frequently in the social sciences, it would at least provide an extra layer of assurance that the results produced are reliable.

4.6 VISUALISATION

With the increasing availability of 'big' data and the ever-complex models being produced, data visualisation has become increasingly important and is in itself a growing field (Liu et al., 2017). Visualisation methods can provide useful ways of presenting the findings of a complex model in a manner that is understandable; for instance, a decision tree is generally much more understandable when presented as a visualisation, rather than a list of rules. Visualisation methods also provide ways of exploring and understanding data sets, which can be crucial in identifying problems within data (such as interactions or outliers), as discussed in section 2.5.1. This section explains t-Distributed Stochastic Neighbor Embedding, as it is a method utilised in the case study chapters.

4.6.1 t-Distributed Stochastic Neighbor Embedding

Newer techniques, such as t-Distributed Stochastic Neighbor Embedding (t-SNE) provide methods of visualising large, high-dimensional datasets on a two-dimensional map (generally, a scatter plot). t-SNE is a nonlinear dimensionality reduction technique and is well-suited for visualising high-dimensional data (Pezzotti et al., 2017). The t-SNE algorithm aims to capture the lower-dimensional relationships within a dataset whilst still preserving the larger global structure of the data; it adapts to the data and can identify different regions and perform different transformations accordingly (Wattenberg et al., 2016). The algorithm has two stages: it constructs a probability distribution such that pairs of similar high-dimensional objects have a high probability of being picked together, whereas dissimilar objects have an extremely low probability of being picked together; and a similar probability distribution is constructed for the low-dimensional objects in the data (Van Der Maaten and Hinton, 2008). The algorithm minimises the Kullback-Leibler divergence between the two distributions in order to create the map. It is essentially a gradient descent problem, and therefore can require many iterations to produce a stable solution, and will produce a slightly different map each time (unless the random seed is assigned a fixed value).

The algorithm has a parameter, called the 'perplexity', which is tuneable and essentially provides a balance between the higher-dimensional and lower-dimensional aspects of the data; it is similar to considering the number of nearest neighbours a data object has (Wattenberg et al., 2016). Another parameter, referred to as 'eta' or 'epsilon' controls

the learning rate. Generally, the algorithm might require many runs to identify the ideal parameters.

A downside of t-SNE is that the plots produced can be difficult to interpret. Wattenberg et al. (2016) points out that where utilised to visualise clusters in the data, it may even-out cluster sizes, so that the actual size of a cluster of objects may be difficult to determine, and the distance between different clusters may not be representative of the actual data. However, Van Der Maaten and Hinton (2008) found t-SNE to outperform seven other non-parametric visualization techniques (such as Sammon mapping and Isomap) when utilised to display data with known clusters (the cluster assignment was used only to analyse the results). Similarly, Platzer (2013) compared t-SNE to Principal Component Analysis (again, using data with known clusters) and found that t-SNE more accurately displayed the structure in the data. Therefore, it would seem that t-SNE may be a useful tool to display clusters in a dataset, however it should be used with caution, and not to make decisions about data.

4.7 Decision Tree Learning

Decision tree learning is a non-parametric, systematic method of predicting a target attribute based upon given predictor attributes. It is a supervised method, as past data is used (i.e. a training data set) to produce a model that may then be used to classify new data. It can be a useful data mining tool as it generally requires little data preparation, produces easy to understand visualisations of the rules produced (if the tree is not too large), and as well as classifying and predicting data it can also be helpful in exploring the structure of a dataset (Rokach and Maimon, 2015). It may also be used as a form of feature selection, by selecting only the useful attributes from high-dimensional datasets (Guyon and Elisseeff, 2003).

Decision tree algorithms generally employ a top-down divide and conquer approach - they start with the whole dataset and then recursively partition the data into smaller and smaller subsets. At each step, the attribute is chosen that best splits the data (this can be a binary or multiple split depending upon the algorithm) with respect to the target. There are various measures of what the 'best' split would be, such as Gini impurity, entropy or information gain, but generally it is the attribute that provides the most homogeneity in the resulting subsets. This process continues (and is applied separately to each subgroup)

until some stopping criterion is reached or it is not possible to split the data any further (Therneau and Atkinson, 2015).

Decision trees are quite flexible and depending upon the algorithm used they can effectively deal with high-dimensional and mixed data (numerical and categorical), missing values, and can be used to predict both categorical (classification tree) and numerical (regression tree) targets (Kuhn, 2013:174; Rokach and Maimon, 2015). They are also considered a 'white box' algorithm, meaning that, in comparison to many other machine learning algorithms where it is not clear how the answer was derived (i.e. 'black box' algorithms, such as neural networks), their rules can be easier to understand and visualise (Kotsiantis, 2013). In order to avoid over-fitting, trees are grown overly large and then pruned down to a more manageable (or understandable) size, or else they are grown to a certain point and then stopped (Loh, 2014).

The non-parametric nature of decision trees means they are particularly suitable for use with complex social science data, which may often contain interactions, non-linearity, non-Normal distributions, and heteroscedasticity (Ritschard, 2014). As discussed in the previous chapters, regression techniques can be sensitive to such data issues and not accounting for them may lead to flawed results. Decision trees, in contrast, do not share the same limitations and can particularly help to identify problems such as interactions. They do, however, have different limitations, which are considered in section 4.7.4.

4.7.1 Decision Tree Example

Decision trees may be plotted horizontally or vertically, and there are various methods of visualising trees in order to make them informative and readable. At the top of a tree is the root node, i.e. the first split, and this produces two (or more, depending upon the particular algorithm) child nodes. Each child node can produce their own children. At some point, the path through the tree ends in a terminal or leaf node, which is where the classification is applied (in the case of a classification tree), or a numerical range or value may be applied, or else the probabilities of having certain values (Rokach and Maimon, 2015:13).

Figure 4 shows an example decision tree plot which predicts the likelihood of passenger survival on the Titanic given various traits (passenger age, sex, class of travel, the number of spouse or siblings aboard, the number of children or parents aboard). The data was provided by the 'rpart' R package (Therneau and Atkinson, 2015) and plotted using the

'rpart.plot' package (Milborrow, 2013). In this case, the root node, is the question of whether the passenger was male; each data record that is male falls to the left branch, and each female record falls to the right branch. At each further split, the data record falls down the left branch if the answer is yes, or down the right if the answer is no. The leaf nodes display the prediction of 'died' or 'survived'. Underneath each class label on the leaf nodes is the accuracy at that node (decimal) and the amount of data that reached the node (percentage). In this example, the accuracy at the 'died' branches is very low.

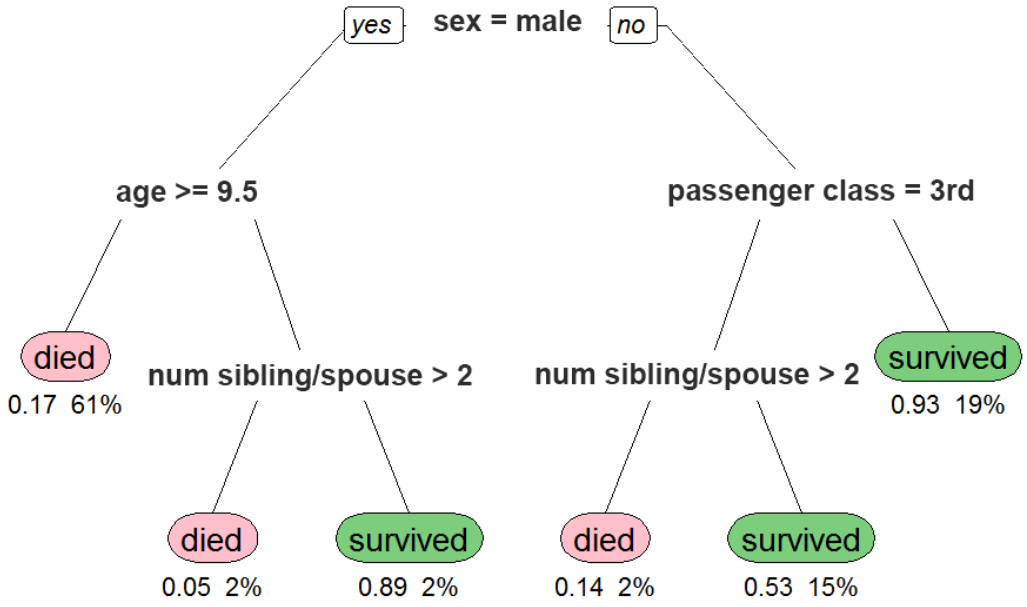


Figure 4: Example decision tree of passenger survival on the Titanic, with accuracy at each leaf (in decimal), and the percentage of data reaching that leaf (percentage). Data obtained from the 'rpart' R package (Therneau and Atkinson, 2015)

4.7.2 History and Development

> Trees, as well as many other powerful data analytic tools (factor analysis, nonmetric scaling, and so forth) were originated by social scientists motivated by the need to cope with actual problems and data (Breiman et al., 1984:viii–ix).

The early development of decision trees grew out of a desire to improve the analysis of social survey data; existing statistical methods were viewed as simply not good enough. Prediction, although useful, was not the primary goal – early research was more broadly aimed at discovering the structure of the data and determining how explanatory attributes were linked to target attributes (Breiman et al., 1984; Ritschard, 2014).

Perhaps the earliest published work was by Belson (1959) who explored the task of choosing relevant predictors and proposed a 'biological classification'. That is, a repeated

(binary) division of a dataset upon different attributes for the purpose of matching two groups for comparison. Belson described his method as a '… movement towards a more empirical way of doing things… [and] a movement away from a sophistication [statistical methods] which is too often either baffling or misleading' (Belson, 1959).

Morgan and Sonquist (1963) published their thoughts on the various problems of handling multivariate survey data, and stated that current methods of analysis were often inadequate in dealing with increasingly complex survey data. Like Belson, they felt that existing statistical methods, and the assumptions they impose in advance upon the data (of linearity, Normality etc.) were too restrictive. Of particular concern was the problem of interaction effects, and the additive assumption commonly used in data analysis, stating '... it is our belief that in human behaviour there are so many interaction effects that we must change our approach to the problem of analysis' (Morgan and Sonquist, 1963).

Sonquist and Morgan (1964) described the first regression tree algorithm, Automatic Interaction Detector (AID), which produced a tree of binary splits that predicted a numeric target, given categorical predictors. It imposed no statistical assumptions upon the data and selected only predictors that were useful to the model. The AID algorithm was then extended to deal with binary categorical targets in THAID (Morgan and Messenger, 1973), and then for multivariate categorical targets in MAID-M (Gillo and Shelly, 1974).

In 1980 a further extension of AID was introduced: the CHAID (Chi-Square Automatic Interaction Detection) algorithm (Kass, 1980), which is still in use today. It used a categorical target and predictors, and employed significance testing (p-values with a Bonferroni correction of the Chi-squared test) to choose the most significant predictor, and could perform multiway (as opposed to binary) splits. In the late 1970s, ID3 (Interactive Dichotomiser) was developed by Quinlan (1986) which utilised information gain as a splitting criterion. An improved version, C4.5, followed this and implemented the ability to deal with missing values, continuous values, and pruning the trees (Quinlan, 1993).

Perhaps the most well-known decision tree algorithm, CART (Classification and Regression Trees) was developed in 1980. Much like the earlier decision tree work, the authors aims were to solve classification and data analysis problems that they felt could not be

adequately solved by existing statistical methods (Breiman et al., 1984). CART added several new features to decision trees: it grew overly large trees and then pruned them back, helping to alleviate previous problems of over and under-fitting the data; it could predict both categorical and numerical targets; and it could handle missing values effectively, by using 'surrogate' splits. CART and C4.5 remain two of the most popular data mining algorithms (Wu et al., 2007).

It should be noted that the focus of much of the work developing decision trees was on providing a non-parametric alternative to methods such as linear regression. The trees were developed to describe interactions and find links between attributes – prediction was generally not the purpose. Their early applications often used survey data and were almost all employed in the social sciences realm. It is striking that their development was prompted by many of the same concerns (about the strict statistical assumptions of the more traditional methods) that still persist today.

4.7.3 CART

An explanation of Classification and Regression Tree (CART) algorithm (Breiman et al., 1984) follows. CART itself is a commercial product not freely available, therefore the 'rpart' R package (Therneau and Atkinson, 2015), which is an implementation of the CART algorithm in the R programming language, was used to build trees for this thesis. It follows the CART algorithm very closely, which is:

1. Start at the root node
2. Search through each attribute to find the one that gives the best split in the data with respect to the target, i.e. it minimises the sum of the two child node impurities
3. Apply step 2 to each child node until some stopping criteria is reached, or else no data remains
4. Prune the tree using cross-validation

Rpart recursively splits the data based on only one attribute, using a binary split at each step, and utilises the Gini splitting criterion for classification trees. The single attribute which 'best' splits the data into two groups is chosen. The data is split and this process is applied separately to each sub-group, recursively, until the subgroups reach a minimum size or else no further improvement can be made (Therneau and Atkinson, 2015). In terms of choosing the 'best' split, Gini attempts to separate data so that each node is 'pure'. The impurity of a node is zero if the node is all one class, and at a maximum if it is

equally divided between the classes. At each step, when selecting where to split, the algorithm selects the split that most decreases the Gini index.

The impurity at each node, A, is defined as:

$$I(A) = \sum_{i=1}^{C} f(p_{iA}) \tag{4.1}$$

Where p_{iA} is the proportion of records in node A that belong to class i for future records, C is the number of classes, and f is the impurity function (Therneau and Atkinson, 2015). The Gini index is generally utilised as the impurity function, and this is defined as (James et al., 2013:312):

$$Gini\ index = \sum_{i=1}^{C} p(1-p) \tag{4.2}$$

Where p is the proportion of records that belong to class i. Using the Gini index equation, if all records in a node were of the same class (p = 1) then the Gini index would be equal to zero, i.e. the node would be pure. The attribute that offers the greatest reduction in node impurity is chosen as the split point (Therneau and Atkinson, 2015).

For regression trees, the Gini splitting criterion is not utilised, instead the split that results in the greatest change in deviance is chosen. The splitting criteria is $SS_T - (SS_L + SS_R)$, where SS_T is the sum of squares at that node (the sum of the expected minus the predicted values squared) and SS_L and SS_R are the sum of squares for the prospective left and right children of that node. This is equivalent to choosing the split that would maximise the between group sum of squares (or minimise the sum of squared error) (Therneau and Atkinson, 2015).

For classification and regression trees, the data ordering is unimportant – all possible splits and all possible attributes to split on are considered at each step (Wu et al., 2007). Whilst, this is a very thorough approach, when dealing with very large, high-dimensional datasets it may lead to performance issues.

4.7.3.1 Pruning

The CART algorithm (Breiman et al., 1984) originated the idea of pruning. That is, to grow an overly large tree, and then prune it to the size that has the lowest cross-validated error; this removes branches of the tree that do not add anything to the model, and

minimises the risk of over or underfitting the data (Loh, 2014). Another reason for pruning a tree is the concept of 'trading accuracy for simplicity' (Bohanec and Bratko, 1994). That is, whilst a larger tree may be more accurate it may be too complex to understand, whereas a smaller tree may be less accurate but easier to understand; therefore, depending upon the purpose, it may be useful to accept a reduction in accuracy in order to produce a smaller, more understandable, decision tree.

The Rpart algorithm defines the risk for each tree as:

$$R_\alpha(T) = R(T) + \alpha|T| \tag{4.3}$$

Where $R(T)$ is the training sample cost of the tree, $|T|$ is the number of terminal nodes, and α is a penalty imposed upon each node (Wu et al., 2007; Therneau and Atkinson, 2015). α measures the 'cost' of adding another variable to the model, and is also referred to as the complexity parameter (CP). An α or CP value of 0 builds a complete tree, and a CP of 1 would build a tree with no splits. α (or CP) is progressively increased from 0 to the value where all splits are pruned away. This is because as α increases the cost-complexity, the tree becomes smaller as the splits at the bottom of the tree that reduce $R(T)$ the least are cut away (Wu et al., 2007). Cross-validation (usually 10-fold) is utilised to find the best value for α; the optimal pruned tree is the one that achieves the smallest risk. In practice the '1-SE' rule is often utilised; Rpart calculates the risk and its standard error during cross-validation and any risk within one standard error of the minimum achieved is considered equivalent to the minimum (Therneau and Atkinson, 2015). This is because, when plotted, the risk tends to display a sharp drop followed by a plateau; therefore the simplest model amongst those 'tied' on the plateau should be chosen (Therneau and Atkinson, 2015).

4.7.3.2 Surrogates Attributes

A key feature of the CART algorithm is that it can handle missing values – many machine learning algorithms either discard records with missing values or would be required to use a method of imputation to deal with them. With CART, records that contain missing values for predictor attributes are retained, and if necessary, surrogate values may be used to account for the missing attribute instead. However, any records with missing target attributes are removed during training, as it is impossible to evaluate a prediction for a record that has no 'answer'; and any record that has all of its predictor attributes missing would also be removed.

At each node, once the primary splitting attribute and split point have been decided, other 'surrogate' attributes are also identified, and these can be utilised should a data record contain a missing value for that split (Therneau and Atkinson, 2015). Surrogate splits are identified even where the data contains no missing values; this means they can be utilised should new data be applied that does contain missing values (Wu et al., 2007). Ideally, a surrogate will split the data identically (or very closely) to how the primary splitter would have split it. Once the primary split has been chosen, the surrogates are identified by re-applying the partitioning algorithm at each split to predict the split using the other remaining predictor attributes (Therneau and Atkinson, 2015). For example, if a split was age < 50 or age >=50, the algorithm searches for any other attributes that might provide a similar split in the data. Surrogates are ranked in terms of performance and up to five are identified by default; only those whose utility is greater than the blind rule of simply selecting the majority class are selected. On the occasion that a record might be missing the primary splitting attribute, and all five surrogate attribute values, then the blind rule is used and the majority class is selected at that node (Therneau and Atkinson, 2015).

Surrogates can also be useful in detecting masking – that is, when two (or more) attributes may be highly correlated in a data set, and so one attribute may obscure the importance of the other/s (Loh, 2014). For instance, consider the two attributes mother's age and father's age – in general, they are likely to be similar and therefore only one might be selected as a primary splitter, but this does not mean that the other is unimportant. To counter this, a Variable Importance score is calculated for all attributes by calculating the sum of the goodness of split for each attribute in its role as either a primary or a surrogate splitter (Therneau and Atkinson, 2015). The inclusion of surrogates can help to reveal any masking, as an attribute which was not a primary splitter in the tree may still have a high importance score (Wu et al., 2007). The rpart algorithm scales the variable importance scores to sum to 100, with the highest score indicating the most important attribute/s.

Aside from the CP value, there are various parameters that may be set to control the tree building process using Rpart – for example, it is possible to specify the minimum number of records at any terminal node, the minimum number at any node for a split to be

attempted, the number of surrogate splits, the maximum depth of the tree, and to weight outcomes using prior probabilities (Therneau and Atkinson, 2015).

4.7.4 Advantages and Disadvantages of CART Decision Trees

As with all machine learning methods, there are both positive and negative aspects associated with the use of decision trees, and this section summarises these aspects. Firstly, the more positive aspects:

Feature Selection: Decision trees can deal with high-dimensional data and perform feature selection automatically. Where a dataset has many attributes only those most important to the target will be chosen for modelling. The variable importance score lists how important every attribute is to the model. This is particularly useful where it is not clear which attributes are important. However, Strobl et al. (2007) make the point that attribute selection can be biased in favour of those attributes with a greater number of unique values, as they offer more potential split-points; this is a general problem affecting methods that utilise impurity reduction measures, such as the Gini Index or Information gain. Therefore, caution should be applied where using data that has attributes with many values; it may make sense to avoid using categorical attributes with many values as their usage can also lead to overfitting (Hastie et al., 2009:310).

Data Preparation: Very little data preparation is required (King and Resick, 2014): there is no need to scale/normalise data (as it will still be split in exactly the same way); missing values are accepted (surrogates may be used); trees are not sensitive to outliers (since data is split within a range rather than on specific values); and trees can handle mixed datasets well, i.e. they can deal with numeric and categorical data together

Interpretability: Compared to other 'black box' machine learning algorithms, the rules of a decision tree are much easier to understand (Kotsiantis, 2013). However, this does not mean that all decision trees are easily understandable. In general, they are intuitively easy to understand and do not require much explanation; it is simple to visualise a tree and its rules. However, where a tree is very large and complex it may be difficult to visualise or understand the rules.

Non-parametric method: Decision trees are non-parametric and many data distributions may be modelled (Rokach and Maimon, 2015:81). In contrast to traditional regression analysis, data that is non-linear, that has interactions or that is correlated may be used.

Decision trees allow users to identify interactions between predictors without the need to anticipate and specify them in advance (King and Resick, 2014).

The more negative aspects associated with the use of decision trees are:

Instability: Trees can be non-robust, and even a small change in the data can lead to big changes in the tree (Li and Belford, 2002; Domingos, 2012; King and Resick, 2014). This is because even a minor change to the training data may cause a split that was initially inferior to the selected split to become superior; and once a different split is selected, the subtree derived from that node may be very different to the original one (Li and Belford, 2002). This may be of less concern where only predictive performance is considered, however where a tree is used to describe data, it should be considered. Methods such as cross-validation and pruning can help to ensure that a tree is not over-fitted, however Hastie et al (2009:312) make the point that instability is 'the price to be paid for estimating a simple, tree-based structure from the data'.

Time: Since (generally) all possible attributes are considered for each split, with a very large/wide dataset this can lead to computationally long running times.

Binary Splits: Many decision tree algorithms utilise a binary split, and this may mean that where a dataset has a particularly complex structure it may be difficult to capture. However, whilst it is possible, and sometimes useful, to utilise multiway splits (for example, the CHAID algorithm can split the data into more than two groups), it is not always practical; multiway splits can fragment the data too quickly meaning there is insufficient data at the next level down (Hastie et al., 2009:311). Multiway splits can also lead to very complex decision trees; whilst a binary split is fairly easy to understand, splits that have more than two leaves may quickly become very complex.

Rare Values: Decision trees are poor at predicting target attributes that contain rare values (Kotsiantis, 2013; King and Resick, 2014). For example, if a target attribute contains the values 'yes' or 'no', and 99% of the data contain 'yes' values whereas only 1% contain 'no', a decision tree may struggle to pick out the 'no' values, as trees tend to go with the majority of the data and therefore would predict 'yes' (with 99% accuracy). However, a method to overcome the problem of imbalanced datasets is to attach weights to the classes (Kotsiantis, 2013).

Predictive Ability: When compared to ensemble methods (such as random forests), single decision trees generally have worse predictive ability, on average about 10% less predictive accuracy than tree ensembles (Loh, 2014). Therefore, if seeking predictive performance alone, there may be more accurate alternative methods. However, a single tree model has a particular advantage over a model built using ensemble methods; it is generally easy to interpret, whereas an ensemble model may be very difficult to understand (Kotsiantis, 2013; Loh, 2014).

In summary, whilst there are advantages and disadvantages to the use of decision trees, it ultimately depends upon the goal as to whether they might be useful. They can, at the very least, aid in understanding the structure of a dataset. They can provide insight into any interactions that might exist in the data and generally highlight relationships between attributes and indicate important predictors. Where predictive power is required, they may not generally be as accurate as ensemble methods, however a decision tree provides an intuitive visualisation that may aid understanding of a problem. In general, it appears that decision trees do not feature very frequently in social science research, and this may be because regression analysis is seen as a more standard method. However, one way that decision trees might be used more frequently is as a complement to regression methods (Thomas and Galambos, 2004; Weerts and Ronca, 2009). Each method can provide a different perspective and utilising the two methods together may provide deeper insight than either method could alone.

4.8 Random Forests, Bagging And Boosting

4.8.1 Bagging

One way of improving the predictive performance of decision trees is via bagging, or bootstrap aggregation (Breiman, 1996). As considered previously, decision trees can be prone to instability (that is, have high variance) and bagging can reduce the variance and produce models that are less prone to over-fitting. Although bagging is often used with decision tree methods it can be utilised with any algorithm. Bagging works by taking repeated samples (all of the same size) from the training data set to create *n* new training sets; this is done with replacement, meaning that observations are repeated in each training set. Then *n* models are built using each of the *n* bootstrapped training sets. The results are combined and the average of all the predictions (or the most frequent class, in

the case of classification) is the prediction of the model. In general, averaging reduces the variance (Hastie et al., 2009:285).

Bagged models can be evaluated without the need for cross-validation or a separate test dataset. This is because each bagged training set utilises around two-thirds of the available data, leaving the other third of data, which was not utilised in the model building process to be used as a test dataset (James et al., 2013). This data is called the 'out-of-bag' (OOB) data and can be used to calculate the OOB error. For each of the *n* models built, the OOB error is computed, and this is averaged for the whole model to produce an unbiased error estimation. Calculating the OOB error can be useful when datasets are very large and so performing cross-validation may be computationally expensive.

Bagged models can be difficult to interpret (Breiman, 1996), as, in the case of decision tree bagging, there is no individual tree to plot (and the interpretability of the single tree plot is one of the key advantages of decision tree learning). However, like a single decision tree, they can provide information about the importance of each predictor to the model. Variable importance for bagged regression trees is calculated by averaging the total reduction of the residual sum of squares due to splitting on a given predictor over all trees; for classification trees the total amount that the Gini index (or other measure) is decreased due to splitting on a given predictor is averaged over all trees (James et al., 2013).

Whilst, in general, bagging might be utilised with any predictive algorithm, Breiman (1996) makes the point that it performs best in terms of predictive accuracy on methods that are unstable (such as decision trees and neural networks), but it can slightly degrade the performance of more stable methods (such as k-nearest neighbours).

4.8.2 Random Forests

Random forests (Breiman, 2001a) are an ensemble learning method. As with bagging, *n* decision trees are built using *n* bootstrapped training samples, but the random forest method differs substantially in that it builds a collection of de-correlated trees (Hastie et al., 2009:587). Where the algorithm deviates from bagging is that at each split in every tree only a random sample of predictors is considered. Where there are p predictors, typically $\sqrt{p}$ are considered at each split for a classification tree and $p/3$ for a regression tree (Hastie et al., 2009:592). For example, in the classification case, if there were 25

predictors, only 5 randomly chosen predictors could be considered at each split (and these would be a new 5 predictors at each split); the best of these 5 predictors is chosen to be the splitter. This means that the algorithm is forced to consider all predictors and as such the ensemble of trees produced are less correlated than they would be with bagging alone (James et al., 2013). This is because, for example, if there was one strong predictor in a dataset, then each tree built would utilise that as the main splitter, and all the resulting bagged trees would be very similar; in this case the model performance may not be much better than simply using a single tree. However, if the model is forced to consider other splitters, it should produce trees that are less correlated and so reduce the variance in the model, producing more reliable results.

As with bagging, random forests are evaluated by calculating the OOB error, and also produce an overall variable importance score in the same way. This variable importance score is useful for determining the most important predictors in a dataset. The two tuning parameters that must be specified are the number of predictors considered at each split (as above, whilst there are typical values, these can be changed), and the number of trees in the forest. Breiman (2001a) proved that random forests do not overfit, however utilising too many (perhaps many thousand) trees may be computationally expensive. Compared to bagging, random forests are generally more computationally efficient, because the algorithm evaluates only a fraction of the predictors at each split, however many trees may be required for the optimal model (Kuhn and Johnson, 2013:200). In general, the ideal number of trees is chosen by considering a plot of the OOB error rate as the number of trees increases. Random forests generally perform better than bagged trees (in terms of accuracy), however similar to bagging, a downside is that they can be difficult to interpret.

4.8.3 Boosting

Boosting (Freund and Schapire, 1997; Friedman, 2001), like bootstrapping can be utilised with any algorithm, although in this context it is discussed with reference to decision trees. The AdaBoost (Freund and Schapire, 1997) algorithm is perhaps the most well-known bosting algorithm. Boosting works by combining many weak models to produce a very accurate prediction overall (Freund and Schapire, 1997). Boosting does not use a bootstrapped sample, and grows trees sequentially, rather than all at once (James et al., 2013:321). Boosting applies weights to the data; one tree is built initially with each

record weighted equally. The records are then re-weighted so that those that were misclassified have their weights increased and those that were correctly classified have their weights decreased, and a new tree is built. This continues with each iteration. Therefore each iteration concentrates on those records that were misclassified previously, and slowly improves the model (Hastie et al., 2009:339).

There are generally three tuning parameters: the number of trees; the shrinkage parameter which controls the rate at which the model learns; and the depth or number of splits in a tree. Boosting generally utilises small trees, this is because each tree takes into account the previous trees built, and so smaller trees are usually adequate (James et al., 2013). Boosted models can overfit the data, so cross-validation is generally utilised to determine how many iterations the model should perform, and experimentation is sometimes required to determine the ideal parameters. Boosting can lead to dramatic improvements in accuracy over single trees (Yang and Wu, 2006; Hastie et al., 2009). However, as with bagging and random forests they are difficult to interpret but do provide a variable importance ranking.

In summary, ensemble methods such as bagging, random forests and boosting all provide interesting alternatives to the use of a single decision tree. If prediction, or discovery of important predictors is the goal then these methods should be chosen over a single tree, as they generally provide higher predictive accuracy. However, if it is important to understand the resulting model (as might be the case for many social scientists) then a single decision tree may be preferred.

4.9 CLUSTERING

Clustering is the art of grouping data items in such a way that items in the same group (or cluster) are more similar to each other than to items in other groups (James et al., 2013). Cluster analysis is an unsupervised machine learning method, and is often performed as a form of exploratory data analysis in order to discover structure within a dataset (Jain, 2010). Cluster analysis is utilised in many fields, such as astronomy, medicine, physics, marketing, biology, genetics, psychology, archaeology (Kriegel et al., 2009; Everitt et al., 2011). The identification of different groups in data allows for separate analyses upon each of the discovered groups. This may mean that any models fitted to these groups may perform better than an overall 'global' model might have done, because a 'global'

analysis may have missed the various contexts within the groups and be prone to an averaging effect.

What constitutes a cluster, and how similar or dissimilar the clusters are depend upon the clustering objective and the data. There is often no over-riding framework of how to perform cluster analysis, and there are many different clustering algorithms available, this means it can be difficult to know which method is optimal (Jain, 2010). Clustering can be an iterative, or experimental, process. The various algorithms mean it is possible to assign all data points into clusters, or to leave out those that do not fit, or to have overlapping (fuzzy) clusters. And the various methods can have very different definitions as to what constitutes a cluster; methods may consider distance, density or statistical distributions to determine cluster assignment.

Clustering algorithms are generally either partitional or hierarchical. Hierarchical methods recursively find nested clusters, whereas partitional methods find the clusters simultaneously as the data is partitioned (Jain, 2010). Perhaps the two most frequently utilised clustering methods are k-means and hierarchical clustering, which are discussed in the following sections.

4.9.1 K-means Clustering

K-means clustering is a method of partitioning a dataset into a user-specified number (k) of different and non-overlapping clusters. The term 'k-means' was first used by Macqueen (1967), but the method was developed by several different researchers dating from 1957 (Wu et al., 2007).

The k-means algorithm requires the number of clusters to be known beforehand – this may be determined either through expert knowledge and/or analysis of the data. The first step in the algorithm is to pick the initial k 'centroids' or cluster representatives; this may be performed by random sampling from the dataset, or other methods such as random partition, or pre-clustering on a subset of data. The algorithm then proceeds by alternating between the assignment step and the update step:

- Assignment step: each record within the dataset is assigned to its 'nearest' centroid; usually Euclidean distance is used as a measure of distance
- Update step: once all records are assigned to a cluster, the centre of the cluster is recalculated by taking the mean of all records contained in it

The algorithm alternates between the two steps, until the cluster assignments no longer change (Wu et al., 2007). At each step, every data record is assigned to one (and only one) cluster.

The K-means clustering algorithm has limitations: it can be difficult to choose the optimal value for k (Jain, 2010); and the algorithm can be sensitive to the initial choice of centroid locations, meaning that an optimal solution may not always be found (Wu et al., 2007). However, this problem may be alleviated by performing multiple runs with different starting centroids. It also forces every data point into a cluster which may not always be desirable, and where the data does not fall into well separated spherical patterns K-means may not perform as well as other methods (Jain, 2010). However, despite its limitations, it is the most widely utilised clustering algorithm and it can provide a quick, efficient and interpretable method for clustering numerical data (Wu et al., 2007).

4.9.2 Hierarchical Clustering

An advantage of hierarchical clustering over k-means is that it does not require the number of clusters to be known in advance. Hierarchical clustering attempts to build a hierarchy of clusters, and there are generally two methods (Everitt et al., 2011):

- **Agglomerative**: Each record starts in its own cluster and the most similar pairs of clusters are merged at each step until one large cluster remains. This is a 'bottom up' method
- **Divisive:** All records start in one cluster and are split at each step until each record is in its own cluster. This is a 'top down' method

In order to decide which clusters should be combined (or divided) at each step, a measure of dissimilarity is calculated. The Euclidean distance between points is most commonly calculated where the data is numeric, but generally any suitable distance measure may be used. Where the data is mixed (that is, contains both numeric and categorical attributes) Gower's general coefficient of similarity (J. C. Gower, 1971) may be utilised. The chosen distance measure is then used to calculate a matrix of dissimilarity, with the 'distance' between each pair of data points calculated by using a linkage criterion. The linkage criterion is used to determine the dissimilarity between all possible pairs of clusters/records at each step. Commonly utilised methods are (Everitt et al., 2011):

- **Complete Linkage**: where clusters consists of a set of points, the distance between groups is calculated as the distance between the most distant pair of points
- **Single Linkage**: the distance between groups is calculated as the distance between the closest pair of points
- **Average Linkage:** distance between groups is calculated as the average of distances between all pairs of points
- **Median Linkage:** distance between groups is calculated as the distance between their centroids
- **Ward's criterion**: the two clusters are joined whose combination results in the minimum increase in sum of squares

The hierarchical clustering algorithm, in the agglomerative case, is (James et al., 2013:395):

1. Begin with *n* data objects (which are each considered as a single cluster), and the pairwise dissimilarities between all data objects
2. For $i = n, n-1, n-2, \dots, 2$:
   2.1. Consider all pairwise dissimilarities between clusters, choose the two that are least dissimilar (i.e. most similar) and fuse them together
   2.2. Calculate the new pairwise dissimilarities for the remaining clusters

The algorithm generally runs until there is one large cluster (agglomerative), or else each cluster contains only one data point (divisive). This sequence of clustering assignments may then be visualised by plotting a dendrogram (Figure 5). A dendrogram is a visual and mathematical representation of the complete clustering procedure (Everitt et al., 2011). It plots each merge (or division) in the data, with the root node containing the group of all data points, and each leaf node containing a single data point. Internal nodes each have two child nodes representing the groups that were merged/divided to form it. The height of a node is drawn proportional to the dissimilarity of its two child nodes (Everitt et al., 2011).

The dendrogram can be utilised to determine any patterns in the data and to decide upon the number of clusters. Whilst there are various methods to determine the optimal number of clusters, in practice this may often be decided by considering the dendrogram (James et al., 2013:393). A cut is generally applied horizontally, where the biggest change in height occurs (where the dissimilarity was greatest), to partition the data into clusters.

In Figure 5, for example, a cut could be made horizontally at a height of 11 to create two clusters. Or where there is prior knowledge of, for example, the existence of three clusters in the data, then the cut might be made at a height of 8 to create three clusters.

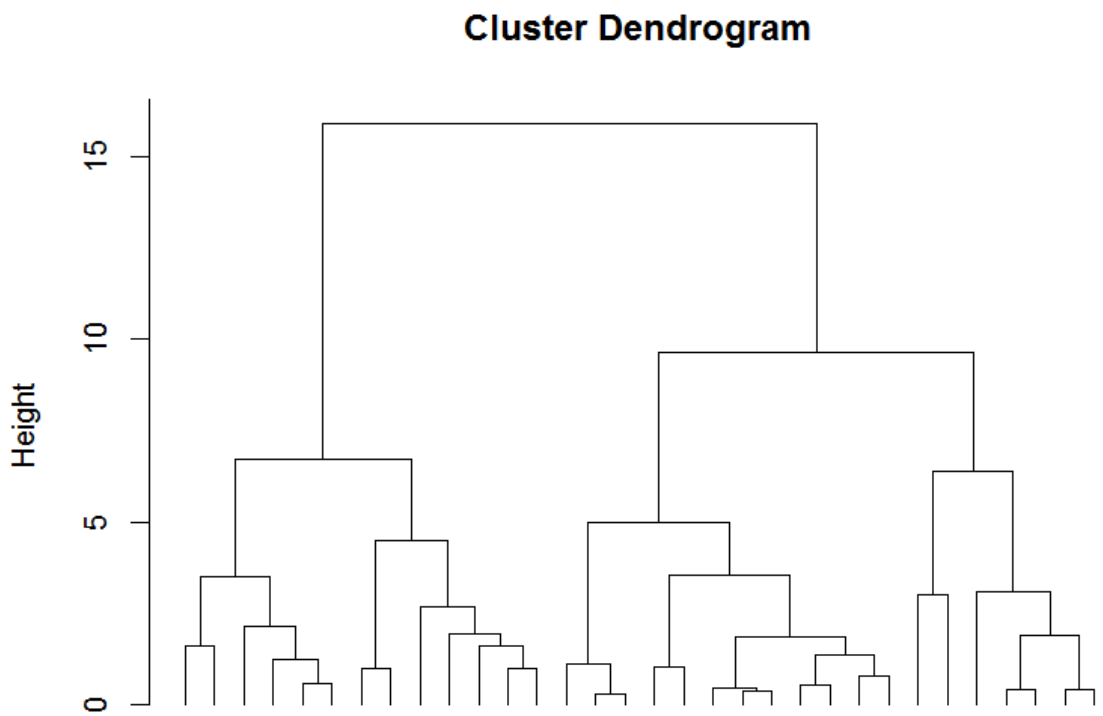


Figure 5: Example dendrogram, plotted using the R base package 'mtcars' sample data

However, whilst useful to describe the structure of the data clustering, dendrograms can be deceptive (Hastie et al., 2009). There are more formal ways of deciding the optimal number of clusters, and as covered in the following section, measures such as the Silhouette value, Calinski and Harabasz index, and Goodman and Kruskal's Gamma coefficient provide alternative methods of determining the optimal number of clusters in a dataset.

A disadvantage of hierarchical clustering is that it does not scale well for large datasets (Everitt et al., 2011:97). For example, a dataset with 100,000 records, would have a 100,000 x 100,000 dissimilarity matrix, which would contain 10 billion data points. Where memory and computing power are an issue, this can make it almost impossible to use for very large datasets. Another disadvantage is that it may impose structure upon a dataset where no structure exists (Everitt et al., 2011; James et al., 2013). The linkage methods have various downsides (Everitt et al., 2011:79): single linkage clusters can be too spread out as they only need one pair of points to be close (resulting in chaining); complete linkage clusters can be too close together, and points within a particular cluster may be

closer to points in other clusters than to points in its own. The average linkage and Ward's method attempt to strike a balance between the single and complete linkage methods. A downside of median-linkage is that it can produce inversions, where a join occurs at a level of similarity that is lower than in the previous step (resulting in dendrograms that cannot be interpreted).

However, perhaps the main advantage of hierarchical clustering is that it does not require the number of clusters to be known in advance (unlike k-means). Another advantage is that any suitable measure of distance may be used; and as a matrix of dissimilarities or distances is computed, the original data is not even required. Also, mixed data might be used; categorical, numerical or binary data would all be suitable provided there was a suitable distance metric.

4.9.3 Evaluation

Given that clustering is often utilised as an unsupervised learning algorithm, there is often no definitive 'answer' as to whether a clustering solution is correct. It is also very difficult to formally define what constitutes a cluster (Jain, 2010; Everitt et al., 2011:7). In general, a cluster might be defined in terms of internal cohesion (homogeneity) and external isolation (separation); however, as Everitt et al. (2011:7) note, whilst there are various definitions 'no single definition is likely to be sufficient for all situations'. Jain (2010:652) states that 'a cluster is a subjective entity that is in the eye of the beholder and whose significance and interpretation requires domain knowledge'.

However, there are various metrics that may help to determine the validity of any clustering. Perhaps the simplest method is examining the clusters and the data in order to determine if they make sense; this may be aided by a domain expert where necessary. If external labels do exist, then the clusters can be compared to those, but this is not usually the case. Clusters obtained by employing different methods may also be compared; similar results might indicate robustness (Everitt et al., 2011:257). Clustering upon subsets of the same dataset and then comparing the results can also indicate whether genuine patterns have been found (Jain, 2010).

To compare two sets of data clusterings (or a data clustering to external labels) the Rand Index may be used (Rand, 1971). This is particularly useful since it can be employed where the number of clusters differ between the two clusterings. It calculates the proportion of data objects that agree; i.e. the proportion that are in the same cluster in

both clusterings, or the proportion that are in different clusters in both clusterings. It takes a value between 0 and 1, where 0 indicates no agreement, and 1 indicates that both clusterings are exactly the same. However, the Rand Index can give large values even where the clusters disagree substantially, therefore the Adjusted Rand Index (Hubert and Arabie, 1985) corrects the Rand Index in order to account for chance. It takes a value between -1 and +1 and provides a more reliable measure of agreement.

One can also measure internal cluster quality or cohesion, and many of these measures can also be used to determine the optimal number of clusters. Milligan and Cooper (1985) performed a detailed comparison of various procedures for determining the number of clusters in a data set, whilst Dimitriadou and Dolnicar (2002) provided a similar comparative study for binary data sets. Milligan and Cooper (1985) found the Calinski and Harabasz index (Calinski and Harabasz, 1974) and the criterion proposed by Duda (1973) to be the most effective for use with continuous data.

The Silhouette Coefficient (Rousseeuw, 1987) is another measure to determine how many clusters are optimal. It is particularly useful in that it can be used for data that is not continuous (many methods are suitable only for continuous data). The silhouette plot provides a useful visualisation and can indicate the quality of a cluster solution; that is, which clusters are well-defined, and which are less clear-cut (Everitt et al., 2011:129). It also provides insight into which individual data objects are well suited to their cluster and which are not. It utilises the dissimilarity matrix and for each data object, i, a value is calculated, $s(i)$, such that $-1 \leq s(i) \leq 1$, where $s(i)$ is the silhouette value (Rousseeuw, 1987).

The silhouette value, $s(i)$, compares the separation of each data object, i, to the heterogeneity of the cluster. Where $s(i)$ is close to 1 it indicates that the data object is well matched to its cluster; that the dissimilarity within i's cluster is much smaller than the smallest between dissimilarity for i's nearest neighbour cluster. An $s(i)$ value close to -1 indicates the opposite; that i is in the wrong cluster and therefore not well matched. An $s(i)$ value close to 0 indicates that it is not clear whether i should be in that cluster or a neighbour cluster. The individual $s(i)$ values may be averaged across the whole cluster to provide a value for the cluster; and across the whole data set to provide an overall measure of the clustering quality. This average silhouette width together with silhouette plots may then be used to determine the optimal number of clusters (Rousseeuw, 1987).

Figure 6 contains the silhouette plot for the hierarchical clustering example that was illustrated in Figure 5. The plot shows the three-cluster solution, as it had the highest overall silhouette value. When viewing the dendrogram alone (Figure 5), two clusters may have been chosen over three; however, the overall silhouette value for the three-cluster solution was higher (0.38 compared to 0.37). Figure 6 shows the three clusters and their individual silhouette values. Cluster 3 is the most cohesive, with a silhouette value of 0.53. The plot indicates that cluster 2 contains two records that may not be ideally suited to it; these are the grey bars that are plotted on the negative side.

In general, silhouette plots of all cluster solutions, and the silhouette values (individually, over each cluster, and over the whole dataset) can be studied in order to decide what the optimal clustering is, and to understand how the clusters fit the data.

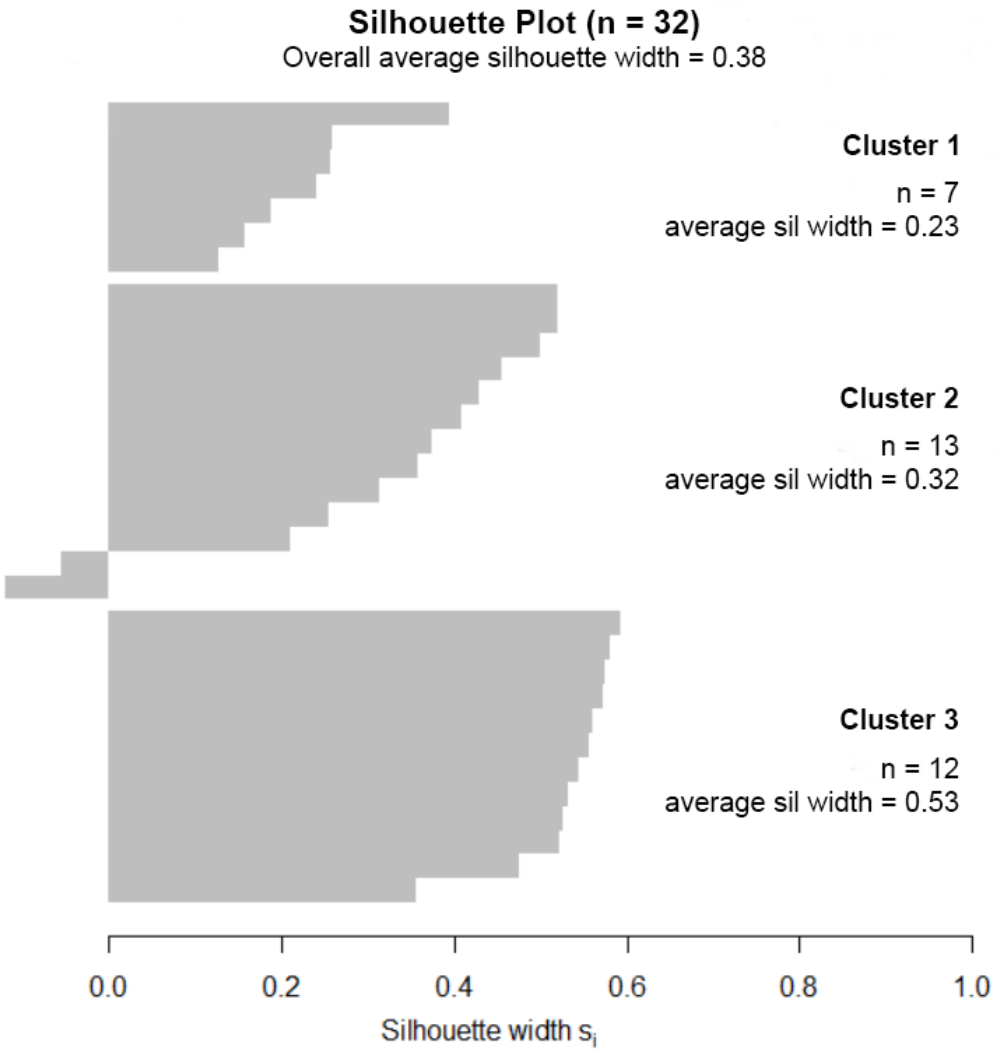


Figure 6: Example silhouette plot, showing the silhouette values for the 3-cluster solution of the example clustering contained in Figure 5, which utilised the R base package 'mtcars' sample data

Another metric which can also be utilised for categorical data is Goodman and Kruskal's Gamma coefficient (Milligan and Cooper, 1985). Like the silhouette coefficient, it utilises the dissimilarity matrix, and compares within group dissimilarity to each between group dissimilarity. It can take values from -1 to +1, and closer to +1 indicates better cluster cohesion.

4.9.4 Limitations

Something which may be considered both negative and positive is that there are so many clustering algorithms and methods available. This means that it may be difficult to determine which method is the most suitable (Jain, 2010). However, clustering is often considered a part of exploratory data analysis, and in this sense, it is normal to consider different methods. Whilst there is no overall framework to follow in terms of choosing which method to use, each method does have its own particular specifications or characteristics, and this can aid in decision making. For instance, k-means is generally suitable only for numerical data (that can be represented in Euclidean space), whereas hierarchical methods can handle numeric, mixed or binary data. However, both of these methods assign all data into clusters, therefore fuzzy clustering methods might be considered if overlapping clusters are required. Often the data and objective of the clustering may dictate the method used.

Perhaps the biggest downside of many clustering algorithms is that they may impose a clustering structure regardless of whether it exists (Jain, 2010; Everitt et al., 2011). This is why, where applicable, visualisation, suitable evaluation metrics and domain knowledge must be considered when evaluating the results of any cluster analysis.

In general, clustering is most effective in lower dimensions; in higher dimensions, many clustering algorithms and traditional distance measurements can become less meaningful (Gan and Wu, 2004; Moise and Sander, 2008; Kriegel et al., 2009; Sembiring et al., 2010). However, it is true that many classical statistical techniques cannot be directly applied to high-dimensional data without modification (Everitt et al., 2011:13). In higher dimensions (i.e. data with many attributes) the volume of the data space increases rapidly, meaning that data points become increasingly sparse which leads to notions of distance and similarity becoming distorted – this is known as 'the curse of dimensionality' (a term accredited to mathematician Richard E. Bellman (Steinbach et al., 2004)). One option to deal with this problem would be to employ feature selection, or dimensionality reduction

techniques, however, clusters generated this way may not fully reflect the original data (Gan and Wu, 2004; Bai et al., 2011). In order to deal with high-dimensional data, traditional clustering methods must be adapted; methods such as subspace clustering, pattern-based clustering, projected clustering and correlation clustering have been developed and this area is an active research field (Kriegel et al., 2009).

In general, numerical data is well suited to clustering; however, clustering categorical data can present problems. This is because, for categorical data there is often no logical measure of distance between points (Gibson et al., 1998; Guha et al., 2000). In smaller data sets with lower dimensions, a similarity or dissimilarity matrix may be computed to compare data points, however this may be computationally expensive in higher dimensions. There are algorithms that attempt to deal with these issues, for example: the k-modes algorithm replaces means of clusters with modes (Huang, 1997); and ROCK which is a hierarchical method that employs links (Guha et al., 2000). However, as with numerical data, clustering high-dimensional categorical data can be a challenge; much of the research into high-dimensional clustering focusses only on numerical data (Bai et al., 2011). Social science research may be affected by this, since social survey data can often be predominantly categorical in nature. However, algorithms are being developed to deal with high-dimensional categorical data (Gan et al., 2006; Bai et al., 2011).

4.9.5 Summary

In summary, this section has highlighted that cluster analysis is an exploratory data analysis technique that aims to discover unique groups in data. The exploratory nature means that there are many different methods, and no one over-riding technique. This means that clustering results should be evaluated by suitable metrics, domain experts or data analysis (such as visualisation), where appropriate. At its most basic level, the discovery of clusters in data can simply provide insight into the data that might otherwise have been undiscovered, and it may generate hypotheses for further investigation. Another possible use for clusters discovered in data is to aid in regression analysis. As considered in Chapter 2, one cause of regression methods not satisfying assumptions is that there may be hidden groups contained in the data; identifying these groups might lead to better performing models on subsets of the data (Achen, 2005).

4.10 LIMITATIONS OF DATA MINING

This section considers the more negative aspects broadly associated with the use of data mining overall. The following chapter considers more closely the practical use of data mining in the literature and the associated concerns, such as privacy.

Much like with the more established social science methods, such as regression and NHST, where poorly implemented, the results of data mining experiments can be unreliable. Perhaps the biggest mistake that can be made when building models is not utilising cross-validation, or a final test (or out of sample) dataset to check a model's performance upon (Domingos, 2012). Many data mining algorithms have the potential to over-fit, therefore results gained only from a training dataset can be vastly different to those where new data is used (Kuhn and Johnson, 2013:62). As datasets get larger and larger, the potential for identifying spurious relationships and noise in data increases. Utilising some form of model validation on previously unseen data is absolutely paramount.

Another way that data mining can be misused is by the equivalent of p-hacking, that is, repeating an experiment until a useful result is achieved, as considered in section 3.3.1. A researcher might change the data and include or exclude attributes until useful results are obtained. Whilst there is not necessarily a problem with this if all experiments are documented, it can be a problem where previous results are not mentioned. This could be unintentional and can be quite easy to do without even realising when using modern statistical software. However, this in itself is not a problem unique to data mining.

Although there are, as previously mentioned in section 4.3, data mining frameworks (such as KDD and CRISP-DM), there is still no common overall framework or methodology (Fayyad et al., 2003; Yang and Wu, 2006; Džeroski, 2007). Yang and Wu (2006) make the point that data mining research can be seen as too 'ad hoc'. This may stem from the fact that data mining is a very broad area, integrating a diverse range of techniques from many different fields and with differing goals (Wu et al., 2007). Domingos (2012) points out that there are a bewildering variety of algorithms available, with hundreds more published each year. It seems therefore that the field is so large and diverse that one overriding framework may be difficult to achieve; this may also be made more difficult by the competing interests involved (for example, academics may have very different goals, and ethical concerns than businesses might).

Another concern is that whilst data mining methods can produce very high accuracy, the results are not always understandable (Burrell, 2016; Hofman et al., 2017). Ensemble methods and black box algorithms can be very difficult to interpret, and Ribeiro et al (2016) make the point that if a researcher cannot understand the reasons behind a prediction, then they may find it difficult to trust (and therefore use) the model. If high predictive accuracy is the main goal, then a lack of interpretability may not be a problem. However, since much social science research is about understanding complex relationships and mechanisms, a lack of interpretability may be worrying. Hofman et al. (2017) note that predictive accuracy and understanding are not necessarily mutually exclusive; it is possible to achieve close to optimal accuracy whilst still gaining insight. Hindman (2015) and Hofman et al. (2017) suggest that machine learning models might be utilised as a benchmark; to compare against more parsimonious models and indicate what is possible (or not). If a machine learning ensemble achieves a particular benchmark accuracy, and, for example, a regression model cannot get close to that benchmark, it may indicate that there is a problem with the regression model. Equally if the machine learning model cannot achieve any acceptable level of accuracy, this may indicate that the particular problem may not be describable with the available data.

As considered in previous sections, it is not always clear which methods are better, or which algorithm to implement when. Domingos (2012:86) suggests that there is 'a lot of "folk wisdom" that can be hard to come by, but is crucial for success'. This may be considered true of most disciplines, but it is perhaps compounded in data mining since the field is so broad. The fact that there are so many options could also be considered positive, since a researcher may have to actively consider which method is best for the particular problem and the data, rather than always using the same method. There are many different software packages available and this may limit the choice of algorithms to some extent. Large commercial packages (such as SPSS) may not always have the most recent algorithms, but the ease of use is an advantage, whereas open source software environments (such as R or Python) can provide more choice and flexibility but offer a steeper learning curve and may be intimidating for those without a programming background.

4.11 CONCLUSION

This chapter discussed the history and development of data mining and provided an overview of some of the methods that are available. It explained the methods that are utilised in the Case Study chapters, and provides the background for the following chapter, which explores the usage of data mining in social science research. In particular, cluster analysis and decision tree methods were explored. Both may be utilised to provide insight into the structure of a dataset; cluster analysis may indicate groups in the data, whereas decision tree learning may identify important predictors and interactions.

Decision tree methods were originally developed by social scientists as they felt that more established regression methods were not suitable for the types of problems and the complex, inter-related data that social scientists often dealt with. As considered in Chapter 2, the problems of satisfying the strict statistical assumptions of regression still exist, and both cluster analysis and decision tree learning may be able to provide a useful complement to regression methods – by identifying hidden groups, interactions and important predictors in data.

The importance of cross-validation, or testing models on unseen (out of sample) data, was considered. This avoids overfitting and may increase the reliability of models; it is one of the most important aspects of any data mining project. Yet, as considered in section 3.3.2, model validation in this way is utilised infrequently in social science research; if it were utilised more frequently it may help to identify problems and could provide an extra layer of credibility to research.

The more negative aspects of data mining were considered; in particular, the fact that there are so many algorithms, and no over-riding framework of how to do things may mean that it is difficult to know which method is best. And where implemented poorly, as with any method, the results may not be reliable. Another concern is that whilst some data mining methods can produce very high predictive accuracy, the results are not always understandable. However, in general, the results of single decision trees and cluster analysis can be easier to interpret and therefore these methods may be more useful to social scientists. Overall, where implemented correctly, data mining methods may be able to provide useful explanatory and predictive power.

5 DATA MINING IN SOCIAL SCIENCE RESEARCH

5.1 INTRODUCTION

Building upon the previous chapter which introduced the data mining process and covered some of the basic methods and algorithms employed in data mining, this chapter describes the literature surrounding the existing uses of data mining in social science research. The growing field of Computational Social Science is considered, and also how Big Data in general might affect social science research. Given the available literature, and considering the previous chapters, this chapter concludes with a list of the different ways that machine learning methods might be utilised to enhance social science research.

5.2 COMPUTATIONAL SOCIAL SCIENCE

Data mining methods do not appear to have been widely adopted by the social sciences; there is little evidence of its use in many subject areas (Scime and Murray, 2013; Veltri, 2017; Yarkoni and Westfall, 2017). However, there are some examples of the use of data mining, and perhaps the most obvious is the growing field of Computation Social Science (CSS). CSS is an interdisciplinary field that aims to harness computational techniques to investigate social science problems. It utilises methods such as modelling, simulation, social network analysis, social geographic information systems (GIS) and largescale analysis of social 'big data'. Cioffi-Revilla (2010) cites its beginnings in the 1960s when social scientists began using computers to analyse their data, however the field has really only come into prominence within the past decade.

In 2009 a group of social and computer scientists published a paper which highlighted that the data-driven field of CSS had been slow to emerge (Lazer et al., 2009). They made the point that other fields such as biology and physics had benefitted massively from modern techniques of collecting and analysing massive amounts of data, but that the social sciences so far had not. They also warned that although CSS was already occurring, it was being performed almost exclusively by governmental agencies such as the USA's National Security Agency and large companies such as Google and Facebook – and if academics did not act soon there was a real danger that CSS might become the exclusive domain of private companies. Similarly, Savage and Burrows (2007) also suggested that where once social scientists had been the champion of innovative methods such as

sample surveys and interviews, such methods were becoming dated and unlikely to sustain future research; to remain relevant social scientists must explore this new social data and the methods required to deal with it.

One way to encourage greater data and analytical skills is by building inter-disciplinary research groups (Lazer et al., 2009; Watts, 2013). In the past few years, far more groups of this kind have been formed (Borge-Holthoefer et al., 2016). Social scientists may generally have little experience with large datasets but they do have deep subject knowledge, whereas data scientists have technical capabilities but may have little training in inferring causal effects, therefore there should be an integral role for social scientists in such collaborations (Watts, 2013; Grimmer, 2015). However, despite the rising interest in CSS, Giles (2012) noted in 2012 that there was still a general lack of awareness of the potential of data; little data-driven work was being published in the top social science journals, and computer science conferences that focussed on social issues did not attract many social scientists. Similarly, Watts (Watts, 2013) noted that whilst thousands of CSS papers have been published (on a variety of topics, such as, social networks and financial crises), relatively few are published in the traditional social science journals; the result of this is that CSS has 'effectively evolved in isolation from the rest of social science, largely ignoring much of what social scientists have to say about the same topics, and largely being ignored by them in return'. Watts (2013) makes the point that one of the main challenges for CSS is to ensure engagement between the communities, so that CSS does not simply become a subfield of Computer (or Data) Science but plays a part in asking important social science questions.

The field of CSS is developing rapidly (Borge-Holthoefer et al., 2016), and the publication of articles is increasing, although Heiberger and Riebling (2016) make the point that compared to overall social science research endeavour, these still seem very few. Whilst the development of CSS was perhaps precipitated by the increasing amount of data available, it is not solely about the data (King, 2016), it is about developing methods that will gain new insight from this data. Conte et al (2012:327) state that the 'traditional tools of social sciences would at most scratch the surface … whereas new tools can shed light onto social behaviour from totally different angles'. Similarly, Veltri (2017:5) suggests that this new data allows the introduction of algorithmic and machine learning methods which bear 'considerable potential for the social scientist'.

One example of the utilisation of CSS is in the integration of case-based modelling with complex systems research. Case-based modelling, which traditionally was not computational, is an established social science technique that conducts in-depth, idiographic, comparative analyses of cases and their configurations (Rihoux and Ragin, 2009; Castellani and Rajaram, 2012). More recently, it has been utilised as a computational technique to model complex systems (Castellani and Rajaram, 2012). This stemmed from Byrne's (2009) suggestion that complex systems should be treated as cases since they have similar characteristics. The Sociology and Complexity Science (SACS) Toolkit (Castellani and Hafferty, 2009) is an example of a case-based model that was designed for studying complex systems. It models a complex system as a set of n-dimensional vectors (or cases), which researchers may compare and contrast; these are then condensed and clustered to create a low-dimensional model of a complex system's structure and dynamics over time and/or space (Castellani and Rajaram, 2012). An important aspect of the toolkit is that it clusters the cases; Uprichard (2009) suggests that cluster analysis is itself a case-based method, in that, in general each case (or record) is assigned to a cluster on a case-by-case basis (according to whether the particular case possesses some level of similarity) and that cluster analysis can be a useful method to better understand cases.

5.3 BIG DATA

Big Data is a term that appears to have many meanings and no set definition (Gray et al., 2015; Kitchin and McArdle, 2016), but generally it is data which satisfies the '3Vs', which are: Volume, Velocity and Variety (McAfee and Brynjolfsson, 2012). Big data is generally large and rapidly growing, it may be unstructured and can take many forms (for example, images, signals, emails, logs, etc.). Much like the usage of the term 'data mining', the term 'big data' is often used as a catch-all, and may indicate the data as well as the methods used on it.

Whilst this thesis is focussed more closely on the use of 'smaller' data in social science research, this section is included as many of the methods utilised on 'big' data may also be utilised on 'small' data, therefore many of the challenges and problems associated with bigger data may still be relevant for small data. Also, researchers may work on both 'big' and 'small' data; there may be much overlap of the two. It is possible to combine many small datasets to make 'big' datasets (Gray et al., 2015), and it may often be

desirable to work on small subsets of 'big' data (Welles, 2014; King, 2016). Hindman (2015) suggests that machine learning methods developed for big data may actually provide the biggest gains (in terms of quantifying uncertainty and predictions) when utilised on smaller datasets.

In 2008, Anderson (2008) suggested that 'big data' was changing the way analysis was performed, and that faced with massive data, theory was no longer important. He argued that researchers would no longer need models or hypotheses because such large data could simply be analysed for patterns. This was perhaps a deliberately provocative article which has prompted much discussion (Chang et al., 2014; Cowls and Schroeder, 2015; Mazzocchi, 2015). Chang et al. (2014) state that no matter the amount of data available theory should still be central to research; they suggest an iterative approach where big data might be utilised to suggest new theory which can then be examined. Mazzocchi (2015) suggests that analysis of big data might inform us of an effect, but that the aim of most researchers is still to explore why that effect occurs. In this sense, big data might be seen as an informational tool, rather than explanatory. However, Shah et al (2015) suggest that big data cannot replace or make more traditional methods (such as surveys, lab experiments, content analysis and clinical trials) irrelevant.

Cowls and Schroeder (2015) suggest it would be almost impossible to analyse data without some kind of hypothesis, or theory; that is, whilst one could trawl big data looking for correlations, it would still need to be placed into some kind of context, and this requires theory. Without even the most basic theory how would we know when something is interesting (it may only be interesting in light of previous research, for instance). However, Chang et al (2014) suggest that Anderson (2008) did have a valid point on some level; they use the example of Google Adwords (it matches users to advertisements) which utilises only data and algorithms, and is extremely profitable yet does not require understanding of the underlying theory. It would seem that it may depend upon the goal of big data research as to the degree of importance placed upon theory. Academics, and in particular social scientists, are interested in causality and understanding the underlying mechanisms, therefore theory is particularly important. Whereas, from a commercial perspective, if profitable insights or decisions can be gained from big data then perhaps this may not always necessitate deeper understanding.

One issue to consider when analysing 'big data' is that it is 'found' data; it often was not collected for the specific purpose it is used for and therefore data collection did not follow the strict rules of a statistically designed experiment (McFarland and McFarland, 2015). This means it may contain many biases, and the sheer size of 'big data' can lead researchers to believe it represents the population when it is actually a very biased sample (McFarland and McFarland, 2015). As 'big' as it is, there will still be subpopulations missing from the data, and conversely, sections of the population that are over-represented. For instance, there may be proportionally less data collected from the elderly, the homeless, or those living in poverty throughout the world (without mobile phones or bank accounts, for example), and those individuals who may who choose to deliberately stay offline (e.g. terrorists or criminals). There is also likely to be proportionally more data collected from the younger, internet connected generation. However, this argument could also be applied to smaller survey data, which is itself often biased. But in this case researchers arguably have more experience and awareness of dealing with neglected groups; they know the data is biased. Giles (2012) suggests that since social networks such as Facebook are increasingly obtaining more users, then there is an argument that they are gradually reducing the bias; also, if biases are understood, then results can be adjusted to account for this in the same way as with survey data. Shah et al. (2015) make the point that even though social media data may not represent the entire population it does not mean that the data is without research value in understanding that population.

Another issue to consider is that researchers still need to learn the limits of big data. For instance, how individuals behave online might not be a true reflection of how they think or feel offline – they might lie online (for example, by clicking 'I voted', when they did not) which will complicate any results (Mann, 2016). One could argue that this is also an issue for social surveys, but this is not a direct comparison, since when an individual completes a survey they know it might be used for research, but when an individual, for example, 'likes' something online they do not necessarily consider that their data may be analysed in future. Analysis could also be biased by online 'bots' that, for example, tweet on mass levels; it may be difficult in some cases to distinguish humans from bots (Shah et al., 2015).

The danger of using traditional statistical methods on such large datasets that easily meet traditional sample size requirements, is that they may lead to 'precisely inaccurate results that hide biases in the data but are easily overlooked due to the enhanced significance of the results created by the data size' (McFarland and McFarland, 2015:1). That is, such large data can produce many extremely significant results, but they may simply be a reflection of the sample size and biases in the data. These problems will persist where researchers continue to focus on using traditional statistical methods on big data. McFarland and McFarland (2015) suggest that data mining methods such as clustering may control for some of the biases (by identifying homogeneous groups to perform analysis on) and help improve accuracy of results.

A consideration for academic researchers is that, in terms of ethics, the individuals who the data pertains to are unlikely to be aware of how their data is used. Privacy and ethical concerns may limit what researchers can do, and it is likely that agreements will need to be made between industry and academia to safeguard privacy (Lazer et al., 2009). There have been various breaches of data privacy over the last few years, and researchers have shown it to be possible to identify individuals out of anonymised data; the sheer quantity of data being collected makes identification more likely (Barbaro and Zeller Jr, 2006; Lazer et al., 2009; Gymrek et al., 2013). Kosinski et al (2013) showed it was possible to accurately identify specific traits about people (ethnicity, religious views, sexuality, intelligence) directly from their Facebook likes.

In terms of replication, the use of big data could present problems. Where largescale datasets are analysed it may be impractical for a researcher to download the data and store it locally; analysis is likely to be performed remotely (i.e. on data stored in the cloud). If this data is rapidly changing and growing, then it later may not be possible to link back to specific data used at a specific point in time (Crosas et al., 2015). It is likely therefore, that there will be a need to develop methods to universally identify and archive data, and also techniques for academics to cite specific sections of data.

An advantage of the collection of so much data is that it may be used to prove or disprove theories for where before there was simply not enough data available (Giles, 2012). Big data might be particularly useful in terms of identifying and studying minority groups. A researcher can now focus on very narrow groups of people who otherwise might not have been included in data in the past (but are now because the data sample is so big), or who

would simply have been lost in the noise. Welles (2014:2) suggests that 'by choosing to make Big Data small, we can rectify historical admissions and biases in social science research and build better, more comprehensive, *bigger* understandings of human behaviour'.

Another consideration in terms of data, is that it is not just new data that can be utilised by data mining techniques. There are various projects to digitise historical data, such as, Google's book digitization project, the proceedings of the Old Bailey (dating from 1674), and records of the Atlantic slave trade. Bearman (2015) suggests that, if directed towards answering important questions and used in context such old 'big data' might revolutionise historical social science. Although he cautions that since many social science historians 'accounts rest on narrative sentences', the usefulness for them may still be limited (Bearman, 2015:1).

Data collection used to be expensive and limited (for example, sample surveys and interviews), but newer methods allow the possibility of largescale data collection that may be refined over time depending upon trends and patterns, and the inclusion of data that might add more context, such as location, time, movement, etc. (Giles, 2012; Chang et al., 2014). And all for much less time and expense. Yet, social survey data should still play an integral role in social science research; Gray et al. (2015) argue that where considering long-term attitudinal trends or patterns, the most reliable data is generally derived from national surveys, as big data cannot usually be created retrospectively. Surveys can provide a greater level of detail not necessarily available from big data; in particular, around how people think. The traditional longitudinal datasets can still provide useful information about trends over time and historical processes, and in terms of refining their data, large surveys manage to refine their questions every year and still remain relevant (Gray et al., 2015). For example, surveys such as Understanding Society, the Crime Survey for England and Wales, and The British Social Attitudes Survey are updated regularly.

It is possible to join many smaller datasets to make 'bigger' datasets; Gray et al. (2015) joined longitudinal survey datasets to aggregate statistics (such as, unemployment and inflation rates, and crime data). The data may be linked via time periods and categories of respondent (such as age, ethnicity, income, location, etc.). Whilst data such as this still may not be 'big' in terms of the number of records, it is 'big' in that it is wide (that is, it

may have many, potentially hundreds or thousands of, attributes). It is also possible to deploy more conventional social surveys online, meaning there is (theoretically) no limit to the amount, or type, of respondents (Burrows and Savage, 2014). This may mean that the data is biased in terms of respondents, which would need to be considered, but it also allows for extra detail such as timestamps and location, and potential for further research such as considering how the survey was shared on social media (for example, via the analysis of tweets) (Burrows and Savage, 2014).

Cave (2016) argues that more data is not always better. The nature of big data means that it may contain a lot of noise and redundancy. Therefore, there is still a need for the 'smaller' datasets there were once synonymous with social science. These datasets, where much thought, time and research were devoted to their content, may well still contain more complex and valuable information, depending upon the research question. And many of the methods applied to big data may also be applied to smaller data. Scime and Murray (2013) argue that data mining techniques are well suited to analysing social science data such as surveys, which over the years may have become broader, more complex and with more missing values. Data mining techniques are particularly useful in identifying relationships and reducing dimensionality. Given the amount of money spent on social science surveys and research, social scientists are 'ethically obligated to conduct comprehensive analysis of their data' and data mining is an ideal tool for this (Scime and Murray, 2013:1).

Given the various advantages and disadvantages of 'big' and 'small' data, it would seem that utilising both may well be useful in some cases. Generating patterns and models on one set and using another for context. Most of all, if social scientists do not participate in the world of big data, there is a danger that their voice will no longer be heard; data scientists rather than social scientists will be the ones generating social theory (Watts, 2013; Burrows and Savage, 2014).

5.3.1 Data Brokers

One example of the use of large-scale commercial social data mining and analysis is found in database marketing companies, such as Acxiom and Experian, who have collated massive collections of consumer information. They are often referred to as Data Brokers, and they capture, gather and combine data from multiple sources in order to repackage it and then rent or sell it (Kitchin, 2014). They collect data such as surveys, consumer

purchase information, financial information (such as credit records), voter registration, court records, mobile phone locational data, property data, web browsing records and social media information (Singer, 2012b; Federal Trade Commission, 2014; Kitchin, 2014). This combination of individual sources of data is used to compile detailed profiles of an individual's life. Data is generally repackaged and sold to commercial companies for purposes such as targeted marketing, identity verification, or fraud detection (Federal Trade Commission, 2014). It is also sold between the various data brokering companies. Since this data is not collected directly from the individual, most people are unaware that it is being collected and used by data brokers (Federal Trade Commission, 2014); in general, most people are unaware of the existence of data brokers (Singer, 2012b).

Data brokerage companies are storing and utilising massive collections of information to build ever more complete profiles of individuals. Acxiom claims to have information on approximately 700 million individuals worldwide, and 'over 3000 propensities for nearly every U.S. consumer' (Acxiom Corporation, 2014:8). This data can be used for predictive modelling, or to derive groupings (Federal Trade Commission, 2014; Kitchin, 2014). In the UK, Acxiom assigns individuals to 55 different clusters, combining details such as life-stage, affluence, age and digital activity (Acxiom, 2017). Examples of particular cluster names include 'thrifty pensioners', 'parents under pressure', and 'going places'. However, in terms of clustering techniques, it is not clear how scientific their methods are as there is little technical detail of how the clusters were derived.

There are concerns that decisions made using this data might be utilised in such a way that individuals could be disadvantaged, or discriminated against (Singer, 2012a; Federal Trade Commission, 2014; Steel, 2014). Data held by data brokers is used to provide information for companies who, for example, score credit applications, or evaluate insurance applications. If data brokers make incorrect inferences from their data about sensitive subjects such as an individuals' ethnicity, age, income or likely health conditions, this could potentially have negative consequences. Based on these inferences insurance premiums might be higher, credit could be denied, or a job application refused, for example. Given this, the quality of the data held by data brokerage companies is very important; inaccurate data could have a profound effect upon an individual. Given the overall lack of transparency around data brokers, an individual may not be aware that incorrect information is held about them, and even if they are, there is no clear

instruction on how it might be corrected (Federal Trade Commission, 2014). Data brokers are also largely unregulated (Singer, 2012a). In terms of security, there have been breaches and hacks over the years (Bambauer, 2013; Thielman, 2015), but there is little information on the effect of these. Any breach is worrying given the amount of information held by these companies.

Such large collections of information should be of major interest to social scientists, yet little academic research appears to have focussed on data brokers (Kitchin, 2014). Given the massive amount of diverse information they possess, and the fact that they are classifying, targeting and predicting individuals' behaviour, it is surprising that there has been so little academic interest in their methods. Both from the point of view of providing a critical assessment (for example: are their methods ethical? Is the data stored securely?), to a discussion on the methods used and the possibility of academics accessing this kind of data for their own research. However, it is likely that the main reason for the lack of research is the overall dearth of available information.

Given the lack of information, it is not clear whether it would be possible for academics to access this data, but there would likely be ethical concerns if access was granted: for an academic study, informed consent must generally be provided, and in the case of this type of data, consent is vague (if provided at all). Private companies that hold data for commercial gain may have no motivation to share, and will also have very different ethical concerns to academia. Whatever the reason for the lack of academic research, the danger is that the insights gained from mining this type of data may remain solely in the commercial sector.

5.4 Data Mining in Social Science Research Literature

Whilst there is not a large body of work, there are examples of the use of data mining within social science research. The following section highlights that there are small pockets of work in certain areas, other less connected examples, and a small field called Educational Data Mining which appears to be growing.

5.4.1 Educational Data Mining

There are a few subject areas where data mining has been embraced, and the small but emerging field of Educational Data Mining (EDM) is one such example. Universities and educational institutions have the opportunity to collect massive amounts of data.

Examples include, logging student (and teacher) progress and grades, records of how students interact with online learning environments and social networks, as well as the collection of data pertaining to the management of institutions, such as the allocation of resources and scheduling of classes. This is largescale data and the potential to use data mining techniques to optimise and explore it is huge. EDM aims to explore educational data and to develop methods to better understand students and the environments that they learn in (Romero and Ventura, 2007; Peña-Ayala, 2014).

Weerts and Ronca (2009) used classification trees (CART) to help predict donations from alumni, providing an exploration of the use of trees for this purpose, and concluding that, whilst there were limitations, the trees provided an informative description of alumni donors and non-donors. It utilised historical data on donations; how much was donated (if anything), coupled with demographic data and information about the alums experience at the university. The study suggested that levels of donation were related to factors such as household income, religious background, the particular degree that was obtained and how the alum kept in touch with the university. Most importantly, the characteristic which most distinguished between those who were likely to give and those who were not, was the perception of whether the university needed their help. The practical consequence of this was advising the university that any future communication with alumni should clearly articulate why their help was needed. A limitation of the study was that it did not perform well for large donors; the authors noted that they did not utilise a cost matrix as it was outside the scope of the study. However, large donors were rare (accounting for less than 1% of donors overall) and, as considered previously, decision trees can struggle with rare values. A cost matrix may have helped, however the authors also considered that it may have made sense to analyse large donors separately. Overall, the decision trees were viewed as much more intuitive and easier to understand than logistic regression, particularly for non-technical users, and the study provided new insight into the motives for donations.

Thomas and Galambos (2004) also explored the use of decision trees (CHAID) to determine factors of student satisfaction, comparing the method to regression (both forward and backward stepwise regression was utilised). Understanding what contributes to student satisfaction is important as it is used to assess the effectiveness of an institution and informs decision making. The data consisted of a student satisfaction

survey that had been administered to a representative sample of undergraduate students. The regression models identified 17 attributes (out of 140) that were strong predictors of satisfaction, and the decision tree broadly identified similar attributes. The authors note that where the decision tree was particularly useful was in identifying different predictors for different groups of students (for instance, students with high intellectual growth had different predictors than those with no growth). However, whilst the CHAID tree illustrated the heterogeneity of the students, the authors note that the results were complex and might be difficult to present to a non-technical audience. Whilst both methods produced broadly similar results, the authors highlighted that each method provided a different perspective, and that using both methods together led to much greater insight than regression alone would have provided.

Further examples of the different uses of EDM include: explorations of the use of decision trees and clustering to explain student success and failure (Salazar et al., 2004); using association rules to identify weaker students for remedial classes (Ma et al., 2000); and comparing different algorithms in order to identify students in danger of dropping out of education (Er, 2012; Manhães et al., 2015). Salazar et al (2004) used a large data sample (over 20000 records) to identify various rules that might indicate good academic performance, and whether a university will retain students. However, some of the methods were vague; the authors mention identifying homogeneous subsets but do not detail how this was performed, and there is also little detail of the clustering which was then performed on those subsets. Ma et al. (2000), state only the size of their test dataset (153 records), therefore it was unclear how large their overall sample was; but the study demonstrated that association rule mining performed better than the current method (a simple threshold) of identifying students likely to fail their A-levels. Identifying these students was important (so that they could receive extra tutoring), however it was costly to identify too many students (who might not all need help), yet important to identify those who genuinely needed help. Both Er (2012) and Manhães et al. (2015) utilised relatively small datasets (respectvely, consisting of 625 and 402 records) that consisted of only time-varying data (such as grades and attendance) in an attempt to identify students who were likely to fail academically. Both studies compared various machine learning methods: Manhães et al. (2015) noted that the less interpretable methods (random forests, neural networks, etc.) had the highest accuracy, but found

Naïve Bayes to be particularly useful because it was interpretable yet still had high accuracy; Er (2012) found the instance-based algorithm K-Star to be most accurate.

Overall, it appears that there is a lack of quality research in some of the EDM literature to date; Johnson (2012) states that there are major methodological flaws in some studies, such as not using test data sets for validation, and reporting vague accuracy rates. Johnson (2012) also notes that some studies are quick to attribute causation, for example Delavari et al. (2008) suggest that a lecturer's marital status had an effect upon their students' performance; this may well be spurious. However, whilst cautioning against the risks of EDM, such as ethical issues and uncritically adopting data mining results, Johnson (2012) acknowledges that there may be overall advantages when performed properly.

Whilst there is a small core of EDM literature, what is striking is that many articles seem to focus solely on the machine learning element. That is, they approach the problems from what might be seen as the technical angle, for example, by considering which algorithm is more accurate, or which method might perform better, see (Kotsiantis et al., 2004; Agarwal et al., 2012; Er, 2012; Acharya and Sinha, 2014; Manhães et al., 2015). The data-driven nature of these studies means that there is little focus on any underlying social mechanisms. However, Peña-Ayala (2014) make the point that many EDM researchers are data miners, and it would seem that much of the EDM literature to date is performed by data miners who are exploring new educational data, rather than social scientists who have adopted data mining methods in order to help explain educational social phenomena. As considered in the previous section, it is possible that deeper insight might be gained from this data if social scientists, who may have more experience of causal effects, were also involved in analysis.

5.4.2 Other Research Literature

There is not a coherent or large body of social science research that utilises data mining methods, but there are examples of their usage. In particular, the use of decision trees and random forests, both for exploratory data analysis and prediction, have proven valuable.

Gutierrez and Leroy (2009), Murray et al. (2009), and Chen et al. (2010) employed decision tree learning to identify important predictors and provide an alternative to regression methods. Gutierrez and Leroy (2009) used data from the USA's National Crime Victimization Survey (38494 records) to identify predictors that influence whether or not

a crime is reported. They argued that traditional statistical techniques (such as regression) result in a limited view of data, and that decision trees which can deal with interactions and complex data, allow the inclusion of a wider variety of predictors, and can therefore provide new insight. Their decision trees were more accurate than baseline accuracy (by 10%) and more accurate than a tree built utilising only those predictors recommended in the literature (by 3%). Their research identified several factors in reporting crime (pertaining to time lost and medical care) that had not previously been investigated and therefore warranted further research. However, they noted that a shortcoming of their approach was that utilising many attributes meant that the decision trees could quickly become complex and difficult to interpret.

Chen et al. (2010) utilised decision trees (C5.0) on a longitudinal dataset (206 records, which followed children from birth and measured development and behaviour) in order to identify the risk factors of parenting stress. The study identified different groups of parents and different predictors for the varying levels of stress. For instance, the main risk factor identified for parents in the high stress group was child development, whereas for those with lower stress levels the main risk factor was child behaviour. The authors also performed analysis using regression, but found the decision tree analysis more informative. However, the study did not mention any model validation, and the dataset was very small, so it may be interesting to know how well these results would generalise to other data. Murray et al. (2009) utilised data from the American National Election Studies (5757 records) survey, and CHAID decision trees to determine survey questions that indicate those likely to vote in presidential elections. The study utilised domain experts to identify appropriate predictors for the model. Whilst the results were somewhat obvious (the top predictor was a survey question on whether an individual intended to vote), the model also indicated that demographic attributes were not useful predictors, which was in contrast to other research literature.

Ruger et al. (2004) utilised decision trees to provide insight into the conceptualization of Supreme Court decision making. Prior to every case argued during the 2002 term the authors obtained independent predictions from legal specialists and used a decision tree model to also make predictions of the outcomes. The decision tree model was trained on data from all 628 cases decided by the court prior to the October 2002 term (there was no overlap with the test data). The study compared the performance of the decision

trees with those of the legal experts in predicting US Supreme Court decisions. The authors expected the legal experts to be more successful and did not believe that a model which did not take into account legal text or doctrine could outperform legal experts. However, overall the model predicted 75% of the cases correctly, whereas the experts collectively predicted 59% correctly. Despite disregarding specific law and the facts of each case, the model was more accurate; it successfully identified patterns that humans did not recognise. The legal experts were more successful than the model in predicting specific areas of law, and also in predicting certain individual justice's votes. The authors found that by comparing predictive performances, overall insight was provided into both the data and the differing processes of human and machine decision making.

Berk and Bleich (2013) and Muchlinski et al. (2016) compare random forests to logistic regression and found random forests to outperform the regression models in terms of forecasting. Muchlinski et al. (2016) argue that, compared to logistic regression, random forests are more able to accurately predict rare events (in this case, civil wars). The study utilised data (7141 records) dating from 1945 to 2000 which measured annually for each country whether a civil war onset occurred, together with various attributes pertaining to subjects such as economic performance, demographics, geography and political situation. The predictive performance of three different types of logistic regression model (classic, Firth rare events and L1-regularized) and random forests were compared. All three logistic regression models failed to predict any civil war onset in the out of sample data; the random forest model correctly predicted 9 out of 20. Interestingly, some of the attributes considered in the literature to strongly influence civil war onset (for example, anocracy and democracy) had little predictive power, when considering the variable importance scores of the random forest. The authors note that the unbalanced, non-linear nature of the data meant that random forests performed better than the regression models. Berk and Bleich (2013) also re-iterate this point – given a complex, perhaps unbalanced, and non-linear decision boundary random forests generally have superior predictive performance; however, where the true decision boundary is simple, methods such as logistic regression compare favourably.

Berk et al. (2016) utilised random forests in order to predict whether or not to release domestic violence offenders awaiting court appearances. The dataset consisted of 28646 records detailing arraignment cases (from a large metropolitan area in the USA for the

years 2007-2011) where an offender faced domestic violence charges. If the individuals were not incarcerated, post-arraignment, one of three outcomes were possible within two years: a domestic violence arrest associated with a physical injury; a domestic violence arrest not associated with a physical injury; or no arrests for domestic violence. Under current practice around 20% of those released after an arraignment for domestic violence were arrested for a new domestic violence offense within two years. The authors suggest their model would cut failure rates (i.e. releasing offenders that go on to re-offend) by half, that is down to 10% (and this was shown on out of sample data). The model incorporated asymmetric costs to ensure that, where the model forecasted an offender to be a good risk, this was based on strong statistical evidence. One issue with any model such as this, is that not all crimes are reported, however one can only work with the available data. The authors also note that an area worth exploring is whether the model might generalise for individuals who were detained until the next court appearance, or who were incarcerated after arraignment; they were not included in the study as they had no opportunity to reoffend. It might be interesting to know if a model could identify individuals who were less likely to re-offend, and therefore for whom alternatives to incarceration might be considered (in order to save money without compromising safety).

Keely and Tan (2008), Kuroki (2015) and Piscopo et al. (2015) utilised random forests to identify important predictors. Kuroki (2015) determined important risk factors for suicidal behaviour among Filipino Americans, using data from the Filipino American Community Epidemiological Study (624 records) conducted in 1998-1999. 87 individuals were suicide ideators, and 39% of these had attempted suicide. Where predicting suicide ideation the model found the most important predictors to be the presence of depressive disorders and substance use disorders. However, these had little predictive power for predicting suicide attempts; the most important predictors in this case were number of family members and family conflict. The authors note that they chose the random forest method in order to examine relationships in the data without the need for formulating hypotheses a priori. Piscopo et al. (2015) used UK census 2011 and governmental data to predict sense of community and participation in English local authorities, and identified predictors which had not previously been considered in the literature, suggesting further research.

Keely and Tan (2008) utilised data from the General Social Survey for the years 1978-2000 to predict various attributes related to income distribution (such as, whether respondents think the government should reduce the income differences between the rich and poor). The study used tree-based methods (CART and random forests) in order to identify preferences across different identity groups. It found that, in general, views on income redistribution were heterogeneous according to race and other socioeconomic factors; views on welfare were heterogeneous primarily according to race. The authors note that the results raised theoretical challenges, as existing theory did not completely explain them; this implied important areas for future research. The study made little mention of error rates, although it stated they were above 60% for the random forests (and even higher for the CART models); this seems quite high and may itself warrant future research into the generalisability of the results.

Artificial Neural Networks (ANNs) do not appear to have been frequently utilised in social science research; this may be because they are generally more difficult to understand than tree-based methods. However, Grossi et al. (2012) used ANNs to analyse factors associated with well-being. The study utilised data from a representative sample of 1500 Italian citizens who completed surveys on psychological well-being and participation in cultural activities. The particular ANN method that was utilised (AutoCM) allowed a mapping of the associations and strength of connections between attributes. The study found presence of disease and cultural access to be the most important predictors of psychological well-being. The authors argue that the ability of data mining methods to capture non-linear interactions and discover hidden effects (that would otherwise be overlooked by more traditional techniques), provides value.

Stambuk et al. (2007) utilised Kohonen Self-Organising Maps (SOM), Principal Component Analysis (PCA) and hierarchical clustering to analyse religious motivation in Croatia. The study utilised a survey on religious beliefs taken by 473 Croatian citizens, and identified three clusters; the different methods confirmed the existence of the three clusters. Ang and Goh (2013) compared logistic regression to data mining methods (decision trees, ANNs and Support Vector Machines) for predicting juvenile offending. The study data sample consisted of 2899 records, pertaining to Singaporean adolescents, containing biographic data and an attribute detailing whether or not they had previously been charged with a criminal offence (5.8% had). All four methods performed well, with high

accuracy rates (greater than 94%) on test data, although the ANN (97.2%) and decision tree (96.6%) were most accurate. The methods identified risk factors for juvenile offending that were consistent with the theoretical and empirical literature, although there was some disagreement between methods, which would warrant further research. The authors suggest that data mining methods are a useful tool to complement existing statistical methods and are viable for use in forensic psychology.

Census data is widely utilised in social science research, and employing data mining methods on this complex data may have the potential to uncover interesting patterns, yet there appears to be little research in this area. Chang and Shyue (2009), and Chertov and Aleksandrova (2013) perform data mining upon Census data, with each suggesting it is a useful technique. Chang and Shyue (2009) utilised association rule mining, k-means clustering and decision tree learning on Taiwan's 2000 Census data, focusing on three groups (single parent families, elderly people over 65, and aboriginal groups). The authors noted that they were not equipped to interpret the results (as they were not sociologists), and that the data mining methods had been used to identify many associations and patterns, but that domain experts should be utilised in order to interpret the results. Chertov and Aleksandrova (2013) utilised cluster analysis on Californian census 2000 data (610369 records) to identify different groups and the factors which influence the decision to have a baby. However, whilst providing some insight into the data, both studies did not particularly conclude anything that seemed useful. As considered in previous sections, and by Chang and Shyue (2009), it is possible that research such as this might benefit from the inclusion of a social scientist to examine the data and results from a different perspective.

Assi et al. (2012) utilise clustering (the BIRCH algorithm) to determine nine different groupings of people, in terms of well-being and quality of life in Europe. The study employed data from the European Social Survey (Waves 1-4), utilising a sub-sample (of approximately 2436 individuals) for each of the 24 European countries, and clustered on 11 attributes. The nine clusters where characterised by different levels of well-being and quality of life, and when compared across countries the proportion of individuals in particular clusters had much variation (for instance, individuals from Nordic countries had a higher chance of belonging to clusters characterised by high well-being than Eastern European ones did). The authors suggested that where previous studies have been

limited in terms of the dimensionality of data, clustering allowed them to preserve as much information as possible. Jiang et al. (2012) clustered daily patterns of human activities (in the Chicago metropolitan area), identifying different groups of people and reinforcing previous research in the field. The data consisted of 30000 individuals who completed a travel survey detailing their activities by time of day. K-means clustering and principal component analysis was utilised; 8 clusters of people were identified using weekday data (such as 'early-bird workers', and the 'stay-at-home') and 7 clusters of people were identified using weekend data (such as 'afternoon adventurers', and the 'afternoon stay-at-home'). Many of these groups had different demographic breakdowns, and the authors suggest that understanding daily travel patterns may have practical uses such as managing congestion.

More generally, cluster analysis may be applied to aid in the detection of interactions; if there are interactions among attributes, the data may naturally cluster into separate groups around these (Melamed et al., 2013). If clustering is performed, the cluster identification can be added to the overall model as a dummy variable and this can be used in order to identify the interaction effects (Melamed et al., 2013). As considered in Chapter 2, when satisfying regression assumptions, it can be difficult to identify interactions and hidden groups in data. Achen (2005) suggests that first identifying meaningful groups in data may lead to more coherent regression analyses on the individual groups; and McFarland and McFarland (2015:3) also suggest that clustering datasets into homogeneous groups can lead to less skewed analysis on those individual groups, because analysis of an entire dataset can produce results that are swayed by whichever 'mixture of populations is predominant at the time'.

Duncan et al. (2008) and Lalayants et al. (2011) advocate the use of data mining to inform child welfare and child protection. They compiled databases of child welfare information which were utilised for exploratory data analysis purposes (rather than predictive purposes). Lalayants et al. (2011) advocate clinical data mining, which incorporates not only individual clinical records, but also available agency information (such as routinely collected data for monitoring and administration). The authors state that child welfare practitioners found the combined data useful, and they suggest an advantage of a multidisciplinary data mining project such as this is that it allows closer collaboration between researchers and practitioners, and a greater concentration on keeping

consistent and complete data records. Duncan et al. (2008) suggest the advantages of compiling and sharing their data on a public website, was the ability to view patterns and extract valuable knowledge, which may lead to improvements in policy and practice in child welfare services.

More generally, Breiman (2001b), Shmueli (2010), Ward et al. (2010), Hill and Jones (2014) and Hofman (2017) suggest that using predictive methods can enhance social science research, although they caution against focussing solely on prediction; predictive methods should be used together with existing explanatory methods. Indeed, much of the reviewed literature in this section has suggested that data mining methods should be considered alongside the more traditional explanatory methods, not as a direct replacement. Breiman (2001b:204) suggests that the best available solution should be utilised for the particular problem, whether that be predictive or explanatory; considering only traditional social science explanatory methods 'imposes an a priori straight jacket that restricts the ability of statisticians to deal with a wide range of statistical problems'.

One reason that predictive methods are not widely utilised in social science research is that prediction is often considered unscientific (Shmueli, 2010). Yet prediction is standard and uncontroversial in the physical sciences (Hofman et al., 2017). As identified by Shmueli (2010:292), predictive methods have various scientific functions: they can help uncover 'new causal mechanisms and lead to the generation of new hypotheses' (perhaps where employed on large, complex datasets that are difficult to hypothesise); suggest improvements to existing explanatory models (by capturing underlying complex patterns in data); provide a 'reality check' by assessing the distance between theory and practice; provide a simple way to compare competing theories by examining the predictive power of the explanatory model of each; and provide a benchmark level of accuracy. More simply, assessing the predictive accuracy of a model provides a way of quantifying its uncertainty (Hindman, 2015).

As considered in section 3.2.3, attributes identified as statistically significant do not necessarily make good predictors; and explanatory power can also be wrongly conflated with predictive power (Shmueli, 2010; Ward et al., 2010). For instance, Shmueli (2010:304) lists examples of research studies that considered the R^2 value to represent predictive power, when it does not. Greater understanding of predictive methods, and a consideration of both types of methods may lead to better understanding. Whilst

explanatory power does not necessarily imply good predictive power there should be some correlation between the two (Shmueli, 2010; Hindman, 2015). Shmueli (2010:305–6) suggests that, even where prediction is not the goal, studies should report the predictive power of a model alongside its explanatory power; similarly, whilst a predictive model may not require causal explanation to be effective, reporting its relation to theory is important for theory building. Providing this extra information about models, where possible, would allow comparison between differing models and theories, and potentially provide more insight into the modelling process. Hofman et al. (2017:488) make the point that 'prediction and explanation should be viewed as complements, not substitutes, in the pursuit of social scientific knowledge'.

Hindman (2015) suggests that, what is even more fundamental for social scientists than consideration of the various algorithms available, is machine learning's focus on model checking. That is, assessing the performance of a model by cross-validation or a holdout dataset; not just on the data that the model was built on. This focus means that machine learning methods are more robust and thus more likely to replicate than traditional methods (Hindman, 2015).

This section has highlighted that machine learning methods appear to be utilised most frequently in two ways: either as an exploratory method (for example, via clustering or utilising decision trees to identify predictors); or for prediction. A mixture of both may also be applied. In terms of exploratory data analysis, the literature illustrated that cluster analysis can be utilised to identify groups in data, and that decision trees and random forests can also identify groups as well as important (or not) predictors and interactions. This exploratory analysis may aid in informing current models, or in generating new hypotheses or areas of research. Tree-based methods are flexible in that they may be utilised for both explanatory and predictive purposes. In particular decision trees can provide an intuitive model that may be simple to understand for non-technical users; however, decision trees can also be very complex and may not always be easy to interpret. Equally, random forests were shown to outperform (in terms of predictive accuracy) other methods such as logistic regression, particularly for complex, non-linear, or unbalanced data; but can also be difficult to interpret (although variable importance scores provide information).

5.5 WAYS THAT MACHINE LEARNING METHODS MIGHT BE UTILISED IN SOCIAL SCIENCE RESEARCH

Considering the previous sections, overall, there are many reasons that social scientists might seek to utilise machine learning methods:

- Data clustering: The clusters themselves might be of interest, or models may perform better on the separate clusters rather than as one overall model
- Decisions trees and random forests can be used to identify important predictors, interactions, and explore the structure of the data. Decision trees, in particular, are generally easy to interpret, however random forests and ensemble methods are less interpretable
- Dealing with categorical data: much social science data is categorical (particularly social surveys), and attributes that are not numeric can sometimes be difficult to deal with. Variable selection, in particular, can be difficult. However, many machine learning methods can efficiently deal with categorical, high-dimensional data
- Prediction: although prediction is not often the focus of social science research, sometimes predictions are desirable. Predictive models can also be useful in providing insight into data and can provide a benchmark of what is possible. There are many suitable machine learning methods, such as ensemble methods, or Artificial Neural Networks and Random Forests.
- Validation of models: one of the most important elements of machine learning is the absolute necessity of validating models (either by cross-validation or the use of a test data set), and there would appear to be no reason that these methods cannot be adopted by social scientists in order to develop more robust models
- Data mining allows the use of a wide variety of data. There is less need to be concerned with sample sizes and weights etc., and many algorithms are non-parametric, so satisfying strict statistical assumptions is often not required

5.6 CONCLUSION

This chapter considered the usage of data mining (or machine learning) methods in the social sciences. Whilst machine learning methods have been utilised for social science research, the literature reveals that this usage is not widespread, and there are gaps in

the literature. Whilst it is difficult to prove an absence of research, the fact that it was possible to identify only two defined subject areas (computational social science and educational data mining) that utilise these methods would indicate this absence of research.

The literature underlined that by not exploring machine learning and more data-driven methods, social scientists are in danger of producing research that is less relevant to the wider research community. In the future, those responsible for generating social theory and interesting research may be those who are proficient in machine learning techniques (such as computer or data scientists), as they are more equipped to take advantage of new 'big' datasets and utilise the full range of methods that can be applied to them. Yet they may lack the expert knowledge that social scientists have in terms of understanding findings, or in generating relevant research questions; that is why it is vital that social scientists be involved.

However, where machine learning methods have been utilised in social science research, the general consensus was that these methods provided a useful complement to existing, more established, methods such as regression. The non-parametric, flexible nature of machine learning methods means that problems can be considered from a different perspective, without the need to satisfy strict statistical assumptions.

6 CASE STUDY PART A: CLUSTERING TROUBLED FAMILIES

6.1 INTRODUCTION

This chapter focuses on exploring social data that is messy and incomplete, as opposed to social survey data which is generally complete and well coded. In particular, it explores what can be accomplished where some prior domain knowledge is provided. A large database of messy, but somewhat interlinked data, was obtained from an English City Council (ECC) who wished to remain anonymous. The data pertained to families in the City's Troubled Families Programme and covered a wide range of events, such as social services records, details of criminal offences, school records and more. In many cases the scope of the data was the whole city, however some records covered only families or individuals specifically in the Troubled Families Programme. ECC felt that there were specific groups or clusters of families within the data, and that a method of identifying them was required. They also felt that where a family lived might be a factor in the type of problems, or the potential outcome of any treatment received. Identifying these clusters of families might provide a greater understanding of the types of families that exist, and also enable more targeted treatment.

This chapter provides a brief description of the Troubled Families Programme, and the various issues surrounding it; also provided is a description of the database, data preparation and the other sources of data that were utilised. The clustering process is described together with a description of the clusters discovered. Visualisations are provided to highlight various aspects of the clusters and also to provide a geographical picture of where the Troubled Families (TF) lived. Decision tree learning was utilised to provide an understandable description of the cluster rules and place them into a more usable context. Machine learning methods were also utilised in order to determine whether the data pertaining to a family's location was a relevant factor in a family's cluster assignment.

In addition to analysing the data, identifying the different groups of TF that exist within it, and considering the geographical location of the families, this chapter builds the groundwork for the chapter following this. Chapter 7 (Part 2) analyses the outcome for

the families and considers the events that occurred in the year following their introduction to the TF programme in order to determine the possible effect of intervention upon the families.

6.1.1 The Troubled Families Programme

The data for this case study was obtained from an English City Council (ECC) as part of a project to explore the use of data mining techniques on their complex database. In particular, the data focused on the Troubled Families scheme introduced by the UK Government in 2011. Then Prime Minister David Cameron made a speech in December 2011 citing that in the previous year an estimated £9 billion had been spent on just 120,000 families through welfare and state intervention (Cameron, 2011). These families had multiple problems, such as crime, unemployment, anti-social behaviour and school truancy. The Government promised to invest £448 million to turn around the lives of those 120,000 families by the end of Parliament (2015), and therefore reduce the amount of public money spent on them.

One of the core aims of the programme was the desire to shift public expenditure from a reactive model based around responding to acute needs, towards a system of earlier intervention whereby problems might be addressed before they escalate (Day et al., 2016). Rather than families dealing with multiple agencies as previously, under the new scheme each Troubled Family (TF) would be assigned a dedicated key worker who would get to know them personally and coordinate any treatment. Each family's needs would be considered as a whole (as opposed to a group of individuals), and practical support provided using a 'persistent, assertive and challenging approach' (Department for Communities and Local Government, 2012:6). Families would therefore receive specific targeted intervention aimed at their specific problems.

Phase One ran from 2012-2015 and claimed to have 'turned around' 99% of families; Phase Two was launched in 2015, with an extra £920 million allocated to the budget, and aimed to help an additional 400,000 families (Bate, 2016).

The TF programme runs only in England, with each local authority assuming control over their own programme. Some Local Authorities refer to it by names other than the TF Programme. Local authorities were responsible for recruiting their own families, and were initially provided with an indicative number of how many TF were likely to be in

their area (Communities and Local Government, 2012). To be identified as TF, the following criteria were considered:

- Have members involved in crime or anti-social behaviour
- Have children not in school (i.e. persistent unauthorised absence above 15%, or exclusions, or in a Pupil Referral Unit)
- Have an adult on out of work benefits
- Cause high costs to the public purse (i.e. local discretion)

All families who met the first three criteria should automatically have been included. Local discretion was used to filter families who might meet two of the criteria but were still a cause for concern. Local discretion could include families with children on a Child Protection Plan, or members with mental health problems, drug and alcohol misuse, domestic abuse, long-term ill health, persistent police call-outs, or under-18 pregnancies (Communities and Local Government, 2012).

The programme operated on a payment by results system, whereby Local authorities could receive up to £4000 for each family. Local authorities received an initial payment for each family recruited, followed by a further payment once a family had been 'turned around'. Each local authority self-declared their results, with payments issued on the basis of this, although it was stated that there may be a small number of 'spot checks' (Communities and Local Government, 2012).

A family was considered to be turned around where:

- each child had fewer than 3 exclusions, and less than 15% absence; there was a 60% reduction in anti-social behaviour for the whole family in the last 6 months; and the offending rate for all minors was reduced by at least 33% in the last 6 months

or

- At least one adult had moved off out-of-work benefits and into continuous employment in the last 6 months

Phase Two of the programme relaxed and expanded the criteria for a TF; each family had to meet two of the following (Department for Communities and Local Government, 2015):

- Parents or children involved in crime or anti-social behaviour

- Children who have not been attending school regularly
- Children who need help (i.e. subject to a Child Protection Plan or who are identified as in need)
- Adults out of work or at risk of financial exclusion, or young people at risk of worklessness
- Affected by domestic violence and abuse
- Parents or children with a range of health problems

And the criteria for success was also relaxed; a family was turned around where they had either:

- Achieved significant and sustained progress, compared with all their problems at the point of engagement

Or

- An adult in the family had moved into continuous employment and off benefits

6.1.2 Intervention Treatment

In the case of ECC, families could be referred for treatment by multiple services. For instance, a referral might come from a social worker, or perhaps a police officer or teacher who came into contact with the family and thought that they might qualify. Sometimes this initial referral might result in no further action, as the family may not meet the criteria (or they may not want to participate in the programme). Where a family did meet the criteria, there were five main intervention treatment types for TF ([Author withheld], 2017):

- **Assertive Outreach (AO):** works with families whose needs are at risk of becoming complex, attempting to look at the root cause of issues and challenge behavioural patterns. Usually a six-month sequenced programme of support
- **Complex Families Parenting Team (CFPT):** delivers parenting interventions to families with a range of complex needs, usually via weekly classes and lasting between 8 and 20 weeks. Caters for families with children aged between 2 and 16
- **Family Intervention Project (FIP):** aimed at the most challenging families, key workers work intensively with families (visiting them three to four times a week) and implement a bespoke method of multi-agency interventions

- **Families First (FF):** works with families whose children are on the edge of care and attempt to keep them together where safe
- **Family in Need Intervention Service (FINIS):** specifically targets families with Children In Need

Families may be referred for, and receive, more than one type of intervention depending upon their needs, and these might run concurrently or over different time periods. Interventions that were completed were classified as 'Planned Endings', those where interventions did not succeed were classed as 'Unplanned Endings'.

6.1.3 Questions Raised About the TF Programme

Since its inception, there have been various criticisms and questions asked of the TF programme nationally. The overall number of TF identified initially (120,000) was likely to be inaccurate, given that the estimate was based on out of date information and used different metrics to those used for the TF criteria (Levitas, 2012). Indeed, it may have underestimated the actual amount of TF (Full Fact, 2012).

The use of the terminology 'Troubled Family' may also be questionable, given that for example, under Phase Two of the programme a family could qualify simply because an adult with long-term health problems was on out of work benefits. This indicates that some families may not be 'Troubled' at all, or at least not in the way that the Government first implied (Anti-social and criminal behaviour, school absence, etc.). Crossley (2018) notes that many families participating in the programme were not aware that they were labelled as 'Troubled', or that they had been 'turned around'. It is not possible to know therefore if these families perceived that their lives had been 'turned around' (Wills et al., 2017). Another consideration is that labelling the families as 'Troubled' stigmatises them (Shildrick et al., 2016). Hayden and Jenkins (2014) make the point that different terminology is used by other countries in the UK running similar programmes, with their focus concentrating on reducing poverty and supporting complex families, rather than the 'troubled' nature of the families.

The initial estimated cost of £9 billion to the taxpayer has also come under scrutiny with suggestions it may be inaccurate, although with little detail as to how the government exactly calculated the number this is difficult to confirm (Full Fact, 2012; Crossley, 2015). The Government's claim of having 'turned around' 99% of TF in Phase One was viewed with some disbelief. The majority of local authorities reported 100% success levels, yet

social policies rarely have almost perfect results, especially when dealing with people who have such complex needs (Bawden, 2015; Crossley, 2015). Such a high figure may simply be due to setting the threshold for success too low; the practice of letting Local Authorities judge and reward their own performance may also play a part. The use of the terminology 'turned around' may have added to the disbelief around the success rates; it has been criticised as misleading as it implies that a family's long-term social problems have been solved, whereas in reality the phrase was indicative of short-term improvements (House of Commons Committee of Public Accounts, 2016).

In August 2016, there was much press interest around a news story that the final evaluation report into the TF programme, written by independent analysts, had been suppressed by the Government. According to the leaked report, the programme had so far had no measurable impact (Cook, 2016; Swinford, 2016). The report was finally released, almost a year late, in October 2016 with the authors concluding that they 'were unable to find consistent evidence that the programme had any significant or systematic impact' across the range of outcomes (employment, child welfare, school attendance, etc.) (National Institute of Economic and Social Research, 2016:1). However, there were some improvements in the lives of families, but they could not be definitely attributed to the programme. It was also acknowledged that there were some limitations with the evaluation; there were data limitations in some areas, and it may have been too early to realistically evaluate the progress of some families (after 12 to 18 months). Also, the point was made that there was wide variation in how local authorities interpreted the programme; some performed better than others, and the averaging effect of the overall analysis may have hidden this (Day et al., 2016).

However, the report did find some positive outcomes: local authorities were found to have transformed their systems and processes in dealing with these families; there was an improvement in local data management systems due to the auditing requirements of the programme; and there were positive changes in practice for assertive key working. Most notably, the report stated that there was a statistically significant impact on how families felt, with many feeling more positive about their future after participation.

However, justification of the TF programme is not the focus of this case study. As a by-product of the TF programme, detailed databases of information were collected and compiled about the TF (and by extension, those who might qualify as TF). This provided a

unique dataset, rich with information, yet messy. The focus was on providing a data-driven exploration of the data; looking for patterns and clusters within the data and exploring the use of machine learning techniques upon it.

6.2 METHODOLOGY

The aim of this case study was to investigate and identify whether there were any clusters, or similar groups of TF, within the data. Consultation with the ECC indicated that they believed that different types of families existed and that it would be useful to identify them. In doing so, it might provide greater clarity on the types of families that exist, and the different types of problems and needs that might be associated with these groups. It might also enable resources to be directed more purposefully at the identified types of families. The identification of these groups would allow analysis on a group level (as opposed to the global level) which might enable better detection of changes or patterns that were not detected for the whole group (i.e. any averaging effect might be reduced). The Government definition of a TF was also compared to the actual data in order to consider whether the families matched those guidelines. The ECC also felt that geographical location, that is, where a family lived, might have been a factor in determining both whether a family was 'Troubled' and also in the success of any intervention treatment, therefore 'place-based' geographical data was analysed to determine any effect.

Initially, the data obtained from ECC was loaded into a MySQL database where it was cleansed, tidied and linked together into a usable format. Exploratory data analysis was performed to look for any patterns, or problems within the data. There was frequent contact with the ECC to query any issues. Hierarchical clustering was performed to identify any clusters of families within the data. This was performed using the R programming language. Families were clustered on the events (such as offences, social care, school absence, etc.) that occurred in the year prior to their entry into the TF programme; a years' worth of events was thought sufficient to understand the type of issues that a family had before they joined the programme.

Decision tree learning was utilised to derive rules and provide further insight into the cluster assignments. An analysis was performed using GIS (Geographic Information System) to link families to various geographical data and determine whether cluster

assignment had any link to geographical or demographic elements. Machine learning methods, together with regression, were also utilised in order to determine whether the geographical data might have any link to the cluster assignment of each family.

6.2.1 Data Description

The data was obtained from the ECC in March 2016 and was stored in a MySQL database on a secure server in a secure location at Manchester Metropolitan University (MMU). The data was accessible only to individuals approved by MMU and the ECC, under secure conditions, and only for research purposes. The ECC accumulated the data under a series of specific agreements with the various authorities concerned, under the condition that it was used for research purposes and not for any specific intervention decisions. Names and addresses were anonymised by the ECC, however, where available, the post code of each family/individual was retained, so that the data might be analysed from a geographical point of view. In the anonymization process performed by the ECC all Identification codes were changed (every individual and family had a unique ID); most importantly, these changes were consistent within the database so that the data could still be linked together (for example, if unique ID A1234 was changed to B6789 this change was performed for every occurrence of A1234 consistently throughout the database).

The database contained 265 different tables, with each describing different information (e.g. one table for school absence, one for criminal offences, one for personal information, etc.). Each table covered a different timeframe. The majority of the data covered the whole of the population of the city, rather than just those from TF.

All records in a table had a Unique ID, which could be used to link them to records in other tables. For instance, offences could be linked to people, and people could be linked to families. The links were made via one large table which contained all link combinations. Despite there being 265 tables in the database, many were empty or not useful. Table 3 lists details of the most useful tables in the database.

Table 3: Details of useful Information contained in the ECC database

Table	Information	Number of records
Absence	School attendance records for all ECC areas Dating from the start of the 2010/11 term to the end of Term 1 in the 2015/2016 term	1418508
Anti-Social Behaviour (ASB) Legal Actions	Anti-social behaviour legal actions for all ECC areas Dating from 2000 (although the majority date from 2009 onwards) up to Feb 2015. Known to be missing data	1393
Children Missing Education (CME)	Children Missing Education for all ECC areas Dating from May 2007 up to April 2016. Likely to have missing data	1567
Department for Work and Pensions (DWP) Benefits	Benefit claims for individuals who were in TF (or those who were associated with or suspected of qualifying for TF) Claims date from 1984 up to Dec 2015, but data is not historical (i.e. those no longer receiving benefits are excluded)	44273
School Exclusions	School exclusions for all ECC areas Dating from Sept 2009 up to Nov 2015	26363
Free School Meals	Free School Meals claims for all ECC areas Data is not historical; it lists only those receiving on the check date, which ranges from Jan to Dec 2015	12900
Housing Benefit	Housing benefit claims for all ECC areas Data is not historical; it lists only those receiving on the check date, which ranges from Jan to Dec 2015	45358
Intervention Event	Records of Interventions for TF. Contains the date of referral, type and status of intervention Dating from June 2009 up to April 2016	6032
Children In Need (CIN) Event	Children in Need events across all ECC areas Dating from 1983 up to July 2015	111337
Child Protection Plan (CPP) Event	Records of Child Protection Plans for all ECC areas Dating from May 1989 up to Dec 2015	11219
Drug/Alcohol (DA) Event	Drug Alcohol events across all ECC areas, from social care records Dating from Sep 2000 (although the majority date from 2008 onwards) up to Jan 2016	4233
Looked After Children (LAC) Event	Looked After Children records for all ECC areas Dating from 1983 up to Nov 2015	22148
Not in Employment, Education or Training (NEET)	Records for those Not in Education, Employment or Training across all ECC areas Dating from April 2010 up to April 2016	11276
Offence	Criminal offences for all ECC areas Dating from Jan 2010 up to Dec 2015	93871
Person	Contains records for all people in the table, such as date of birth, gender, and various IDs (such as Student and Social care)	617944
Person to Address Via Event	Links people to addresses (post codes) via various events (offences, CIN events, etc.) Dating from 1932 up to 2029 (there were some inconsistencies)	459247
Pupil Referral Unit (PRU)	Records of individuals in Pupil Referral Units Dating from June 2008 up to April 2016. Appears to be incomplete	1337

There were 617944 individual people contained in the database. The quality of the data varied; many tables were obtained from different agencies and some may not have been well or consistently maintained. The database contained duplicate people, that is, two or more records pertaining to the same person. Records identified as duplicate were

merged into one. Duplicates were considered to be records that shared one or more of the same ID (such as a social care ID, or student ID). 2855 duplicate records were removed (merged), leaving 615089 individuals in the database. It was notable that whilst people could be duplicated, events were not; for instance, where two records were merged into one, this did not mean that the events linked to those records were counted twice. There was no evidence of duplicate events within the database; each criminal offence, or school exclusion, for example, was reported only once, regardless of the individual it was attributed to.

There were inconsistencies within the data; dates of birth ranged from 1798 to 2049, implying some mistakes in a small minority of the records. However, the majority appeared to be credible. The data also contained a date of death; 62646 individuals were deceased. However, this data may not have been up to date, and the number of deceased was likely to be higher (given that 4347 individuals in the database were aged 100 or over on the 1st Jan 2016). Table 4 contains a brief description of the individuals contained in the database.

Table 4: Brief description of all individuals contained in the ECC database

Number of people	**615089**
Gender	Female: 46%, Male: 44%, Unknown:11%
Age	Average age = 36 42% of individuals were aged under 25 (calculated on 01/01/2016)
Location	65% could be linked to a postcode
Troubled Families	2% belonged to a TF
Students	16% had a student ID (are or were at school)
Social Care	68% had a social care ID; only 14% link to any social care events
Events	77% linked to no events (offences, school absence/exclusion, social care, etc.)

The other sources of data utilised were:

- OFSTED report data for all schools in the ECC area, obtained from (Department for Education, 2016). Where possible, the OFSTED rating of the school each Troubled Family child attended during the year before first intervention was linked to the child. This was to consider whether there was any pattern related to the types of school that TF children attended
- Office for National Statistics post code data (Office for National Statistics, 2016) which was used to link post codes to various geographical markers (Output Areas,

etc.). Whilst the ECC database contained post codes, in order to link these to other geographical data, markers such as the Output Area classifications were required

- 2011 Census data for the ECC area, obtained from (UK Data Service, 2011) and utilised to link various demographic data to geographical markers (Output Areas, etc.). The Census data was linked in order to consider whether there were any patterns surrounding the characteristics of where a family lived
- Police data for the ECC area, detailing crime and anti-social behaviour incidents linked to geographical markers (Home Office, 2016). Whilst the ECC data contained information on crimes, it consisted of listing individuals who had committed crimes (and the type of crime), but not the location of crimes; the Police data was obtained to provide information on where crimes were occurring

6.2.2 Troubled Families Data

Individuals who were involved in the Troubled Families programme each had a Troubled Family Identification Number (TF ID), which was utilised to identify TF members from the data for analysis. Each unique family had a different TF ID; therefore, members of the same family could easily be grouped together. 13111 individuals had a TF ID, belonging to 4160 unique families. However, the presence of a TF ID did not confirm that a family actually participated in the TF programme or received any intervention treatment. Families were assigned a TF ID when they were referred for treatment, sometimes these referrals were found to be inappropriate (and therefore no treatment was received), or else families may have chosen not to participate in the programme. Therefore, to confirm that a family actually participated in the programme, the record of Intervention Events was consulted. Of the 4160 families with a TF ID, 2555 families (comprising 8447 individuals) actually received any treatment.

In order to cluster on just the events that happened in the year preceding entry into the TF programme, data was compiled that counted all events occurring in the year prior to a family's first intervention date. The first intervention date was taken as a family's date of entry into the TF programme. Where a family had multiple intervention referrals, the first date that resulted in treatment was used. If a family was referred at an earlier date, but did not receive treatment at this time, this date was discounted. It was felt that a

years' worth of data provided a sufficient amount of time to build up a picture of events surrounding that family's life before intervention.

Given the various timeframes of events data in the database, and to ensure that a years' worth of data existed before the first intervention date, only TF with a first intervention date between 1st August 2011 and 31st July 2015 were included for analysis. This allowed a window of four years for analysis and included 2155 families. All events that it was possible to count were counted (for example, school absence, criminal offences, social care events, benefits data, etc.). Since most events linked to the individual, rather than to the family, all events were counted on an individual level and then summed together for each family (where appropriate). For each family, a count was made of how many people were in the family, how many were of each gender and how many were children and adults. An individual aged under 18 on the date of first intervention was counted as a child. Table 5 details the family composition, in terms of how many adults and children there were in each family.

Table 5: Number of troubled families with each configuration of adults and children, using ECC TF data

	Number of adults in family							
Number of children in family ↓	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
0	0	252	77	21	7	1	0	358
1	72	325	209	29	9	3	1	648
2	66	251	193	34	10	0	2	556
3	26	159	128	21	6	0	1	341
4	14	59	53	9	4	1	1	141
5	3	25	26	6	0	0	0	60
6	1	11	17	2	1	0	0	32
7	0	7	7	1	1	0	0	16
8	0	0	0	0	0	0	0	0
9	0	1	0	0	0	0	0	1
10	0	1	0	0	0	0	0	1
Total	182	1091	710	123	38	5	5	

The most common configuration for families was 1 adult and 1 child (325 or 15% of families). 182 (8%) families had no adult (consisted of only children), and of these 72 (40%) consisted of a single child. Conversely, 358 (17%) families had no child (consisted of only adults). The most common configuration of adults was one adult in a family (1091 families, or 51%). Of those single-adult families, 886 (81%) of the adults were female. Table 6 contains an overview of the TF data, together with the percentage of families who had each of the particular events occur in the year prior to first intervention.

Table 6: Overview of TF demographics and events occurring in the year prior to intervention, using ECC data

Number of TF	**2155 Families, comprising 7057 individuals**
Family Size	Range: 1 – 11, average 3 members 320 (15%) families consisted of only one member
Gender	Female: 53.7%, Male: 45.9%, unknown: 0.5% 442 (21%) families were all female; 198 (9%) families all male
School Events	Any school absence: 41% of families Absence greater than 15%: 7% of families Exclusions: 12% of families Children Missing Education: 0.1% of families Pupil Referral Unit: 2% Not in Education, Employment or Training: 4%
Social Care Events	Children in Need: 41% of families Child Protection Plans: 17% of families Looked After Children: 8% of families Drug/Alcohol Events: 2% of families
Criminal Offences	Committed by Adults: 12% of families Committed by Children: 8% of families Classed as domestic abuse: 3% of families
Anti-Social Behaviour Legal Actions	0.5% of families (There was missing data here, so it is unlikely to be accurate)
Benefits	43% of families receiving DWP benefits 49% of families receiving Free School Meals 67% of families receiving Housing Benefit (Benefits data was not historical, therefore may not be accurate)
Derived attributes	8% of families with Domestic Abuse issues 3% of families with Drug and/or Alcohol issues

Whilst Table 6 provides the detail of individual events, in more general terms the following were true:

- **State benefits**: 72% of TF received some form of state benefits (DWP benefits, housing benefit or free school meals, though as will be considered in the following section, this data may not be complete)
- **Child safeguarding**: 50% of TF had child safeguarding events in the year prior to first intervention (Children In Need, Child Protection or Looked After Child events)
- **Education:** 44% of TF had events pertaining to education in the year prior to first intervention (school absence, exclusion, children missing education, children in a pupil referral unit, or members who were not in education, employment or training)
- **Crime/Anti-Social Behaviour**: 17% of TF had members involved in crime or anti-social behaviour in the year prior to first intervention (Offences committed by adults, or children, or anti-social behaviour legal actions)

There were three attributes detailing child safeguarding events: Children In Need (CIN) events are the lowest level and occur where there are concerns about a child and they need some form of help; Child Protection Plans (CPP) occur where there are concerns about the safety of a child; and Looked After Child (LAC) events occur where a child has been placed into social care.

It was possible to extract further information from the various attributes, for instance for each Children in Need (CIN) or Child Protection Plan (CPP) event, there was a subcategory detailing what type of event had occurred (such as domestic abuse, or neglect, etc.). This was used to calculate two extra attributes (as listed in the 'Derived Attributes' section of Table 6): an overall Domestic Abuse attribute was created by combining CIN, CPP and Criminal Offence data; and an overall Drug/Alcohol abuse attribute was created by combining the CIN, CPP and existing Drug/Alcohol data. Whilst they were partial subsets of the attributes they were derived from, and therefore might not have been useful for building clustering models, they were derived to provide extra information for the overall analysis.

The Department for Work and Pensions (DWP) benefits data contained details pertaining to claims for six different types of benefits: Job Seekers Allowance, Employment and Support Allowance, Income Support, Incapacity Benefit, Severe Disablement Allowance and Carers Allowance. These are considered 'out of work' benefits and indicate that the recipients were probably not employed. It was noted that the DWP benefits (and other benefit data) was not necessarily accurate, as not all historical data was retained. In the case of the DWP benefits data, it was accurate where a TF had been receiving benefits at the time of first intervention, and still were currently; however, where a family had been receiving benefits but were no longer, this data had not been retained. An indication of this missing data is provided by the fact that across the whole TF programme in England, 63.8% of families had an adult claiming DWP benefits in the year prior to intervention (Department for Communities and Local Government, 2017); the ECC data found only 43% of families, which was some way off. The missing data was disappointing as it meant that it was impossible to identify families that had presumably had an improvement in their circumstances at some point and moved off benefits. This was also a key criterion of the Government guidelines for a family to be considered 'turned around', therefore, it meant that an accurate analysis of whether the families in the ECC data could be

considered ‘turned around’ according to the Government guidelines could not be performed.

The free school meal and housing benefits data were also considered unreliable, as the data was only accurate on the date it was collected, and therefore not historical. Also, it was noted that from September 2014 onwards all children in the first three years of primary school were eligible to receive free school meals, rendering this data less meaningful. The Anti-Social Behaviour, Children Missing Education (CME), Social Care Drug/Alcohol (DA), and Pupil Referral Unit (PRU) data were acknowledged by ECC to be missing records and were sparse. However, all attributes were included in initial exploratory data analysis.

The data was analysed for correlations, using Pearson correlation, as plotted in Figure 7; a blue colour indicates positive correlations (red indicates negative), and the depth of the colour indicates the strength of correlation. Aside from being in receipt of the different state benefits, there was an absence of strong correlations. This was evidenced by the pale colours in the plot; the majority of attributes had weak correlations, less than 0.1. Receipt of the various benefits was most highly correlated, with receiving free school meals (FSM) and housing benefit (HB) having a correlation of 0.69, and DWP benefits having correlations of 0.32 to FSM and 0.42 to HB. This was perhaps to be expected as receipt of the different state benefits are often linked.

Other correlations that were more notable (but still weak) were: school exclusion and having a family member in a Pupil Referral Unit (0.25); school absence and school exclusion (0.22); and school absence and criminal offences committed by children (0.22). This may reflect that some families with children who miss school may have a higher likelihood of having members involved in youth crime and school exclusion (and by extension attendance at a Pupil Referral Unit (PRU), since children who are excluded from school often go on to attend a PRU). More generally, it has been shown that persistent school absence is correlated with crime, and that a quarter of school-age offenders have persistent school absence (The British Psychological Society, 2017)

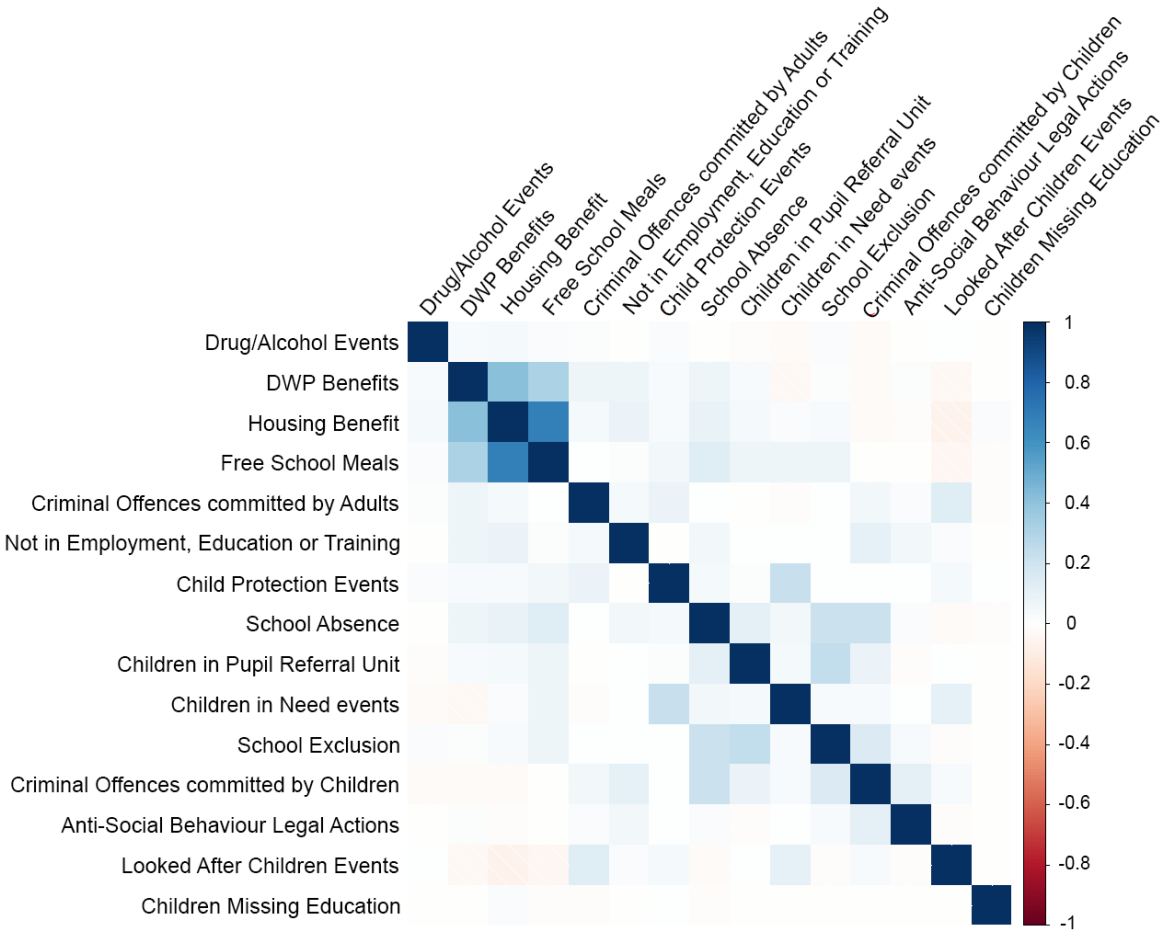


Figure 7: Pearson correlation for various events occurring in the year prior to first intervention, utilising the ECC TF data

The child safeguarding events, CIN and CPP were correlated, as might be expected (0.21). It was notable that, of all the attributes, Looked After Children events (LAC) had the most negative correlations with other attributes. Whilst the correlations were very weak, LAC events were negatively correlated with school absence, school exclusion and being in receipt of the various benefits. This might suggest that families with children in care are perhaps less likely to have school-related issues. LAC events were most highly correlated with criminal offences committed by adults (although this was still low at 0.13), this may cautiously indicate that some families with children in care are more associated with crimes committed by adults. The Government report into the TF programme in England (Ministry of Housing Communities & Local Government, 2018) found that nearly a third of families with LAC events had at least one member of the family who had committed a criminal offence.

For each TF, data was compiled to evaluate the proportion of families receiving each Intervention treatment type (Table 7). The details for each treatment type are contained in section 6.1.2. The CFPT, FIP and AO intervention types were the most commonly

utilised methods of treatment for a first intervention. The FINIS type had a low percentage, but this was because it was a relatively new treatment and had only been utilised for the last year that data was collected. The 'Other' label accounts for one family who received a different type of treatment, that was not detailed in the data.

Table 7: Percentage of TF receiving each first intervention type, from ECC TF intervention data

Intervention Type	Percentage of TF receiving
Complex Families Parenting Team (CFPT)	29.5%
Family Intervention Project (FIP)	27.7%
Assertive Outreach (AO)	24.5%
Families First (FF)	15.2%
Family In Need Intervention Service (FINIS)	3.1%
Other	0.05%

Table 8 details the outcome (or status) of each intervention. Just under three quarters of first interventions resulted in a planned ending, and a fifth had an unplanned ending. An unplanned ending would indicate that the treatment was not completed, perhaps because a family did not want to have further involvement, or if the treatment was not suited to their needs. Interventions were classed as Open when they did not have an end date; this indicated they were still ongoing when the data was collected.

Table 8: Status of first interventions, from ECC TF intervention data

Status of Intervention	Percentage of TF
Planned Ending	74.8%
Unplanned Ending	19.9%
Open	5.3%

6.2.3 Geographical Visualisation of Data

All but nine of the TF (2146 out of 2155) could be linked to a Post Code. Where a TF had multiple addresses, the address that was most recent before the First Intervention Date was utilised; if there was no address before, the address that was dated most recently after the First Intervention date was utilised. The Post Code could then be linked to various geographical markers, such as the Output Area (OA), and Lower Layer Super Output Area (LSOA) codes by utilising the ONS Postcode Directory (Office for National Statistics, 2016). An Output Area is the smallest geographical area for which Census aggregate data is provided and may contain between 40 and 139 households (Office for National Statistics, 2017). The LSOA covers a larger geographical area and generally contains between 4 and 6 Output Areas combined.

Once linked to these codes, each TF could then be linked to aggregated data for their particular area. Demographic data (such as tenure, ethnic group, and education levels) was obtained from the 2011 Census (UK Data Service, 2011). Information detailing individual crime and anti-social behaviour incidents was obtained from data.police.uk website (Home Office, 2016); this contained a monthly list of all incidents handled by the Police force and linked to LSOA. Latitude and Longitude were also supplied; however, these were anonymised to only point to an approximate location, therefore, the LSOA was utilised for location identification. Police data was obtained for the time period covering August 2011 to July 2016, which covers the timeframe of analysis for a TF's first Intervention, with a year added at the end. Although the ECC database already contained information about crimes, it did not contain the location of the crimes; the police data was obtained to provide this information. It was also thought that it might provide contrast to the ECC data, as this provided the location of people who committed crimes, and the police data provided the location of crimes (although there was no way to link the two).

This data was then collated to produce an overall count (and percentage, where appropriate) of the demographic data and Police data for each OA and LSOA. Since there were many categories of Police crime available, an overall count of crime, together with counts of Anti-Social Behaviour, Violent Crime and Burglary were included, as these were the three most populous categories; they were calculated on a yearly basis for each area.

Figure 8 plots the percentage of TF living in each LSOA; that is, the number of TF in a particular LSOA divided by the overall number of households in that LSOA (as given by the 2011 census). There are 282 individual LSOAs in the ECC area. On the date of first intervention, at least one TF lived in 239 of them; 43 (15%) had no TF residing there. The maximum number of TF living in any one LSOA was 34, the median was 7. There appeared to be two main areas where TF lived in higher proportions; one in the North-Western area of the city, and one in the East.

Figure 9 plots the percentage of TF living in each OA and provides a more fine-grained view. There are 1530 individual OAs, and there were TF living in 845 of them on the date of first intervention. 685 (45%) contained no TF. The maximum number of TF living in one OA was 19; the median was 1. Both plots highlight that there was a greater proportion of TF located in certain areas of the city (notably the North-Western and

Eastern areas) and that these areas with a higher density of TF tended to be clustered together.

Figures 8 and 9 contain no identifying geography (the underlying map is not visible) in order to protect the anonymity of the city involved. The range of the bins on the plots were chosen as percentiles; there was one clear group for areas with no TF, then the remaining TF were split evenly into groups (hence different range sizes for the bins).

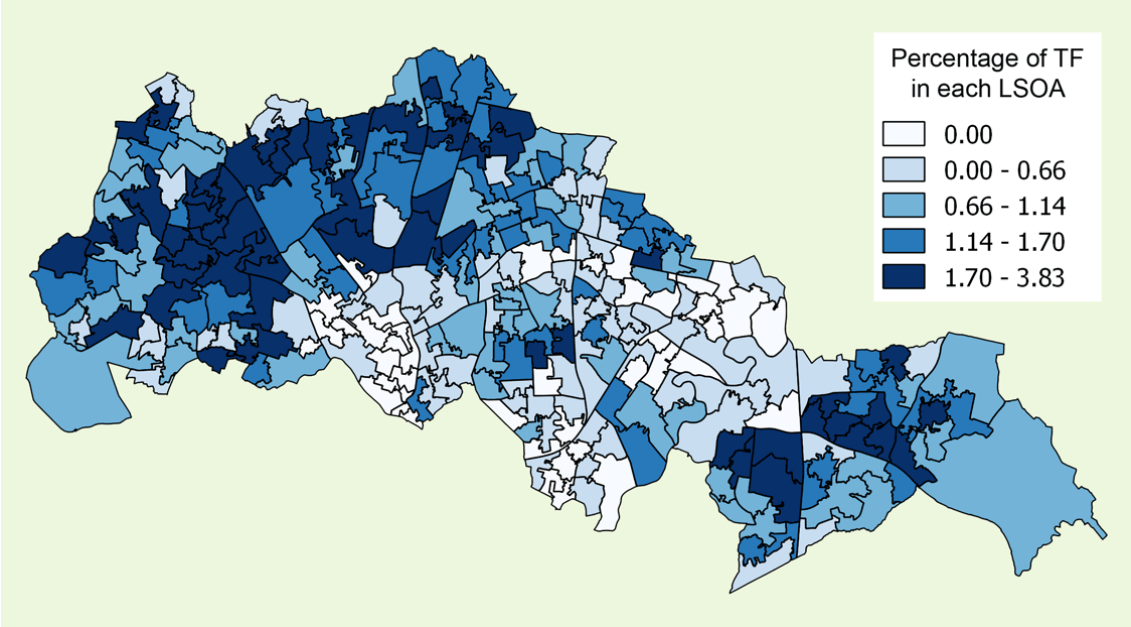


Figure 8: Percentage of Troubled Families living in each LSOA (as a percentage of all families living there). Using ECC data and Census 2011 data

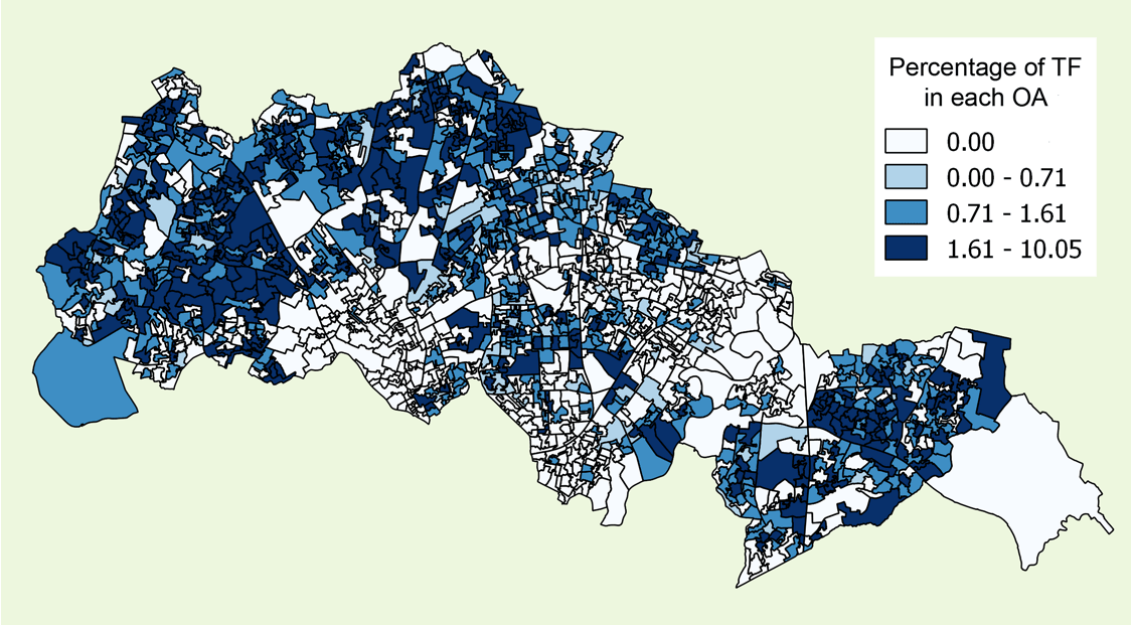


Figure 9: Percentage of Troubled Families living in each OA (as a percentage of all families living there). Using ECC data and Census 2011 data

For each Output Area, demographic data was compiled (from the 2011 Census data); this included data such as the percentage of people with no qualifications living in each OA, and the percentage of lone-parent households, etc. The percentage of TF living in each OA was also included and this was utilised to plot the correlation between the various attributes, Figure 10. Whilst there were some strong correlations between the various census attributes (as evidenced by the dark colours), it was the correlations with the percentage of TF living in an area that was the focus for this analysis.

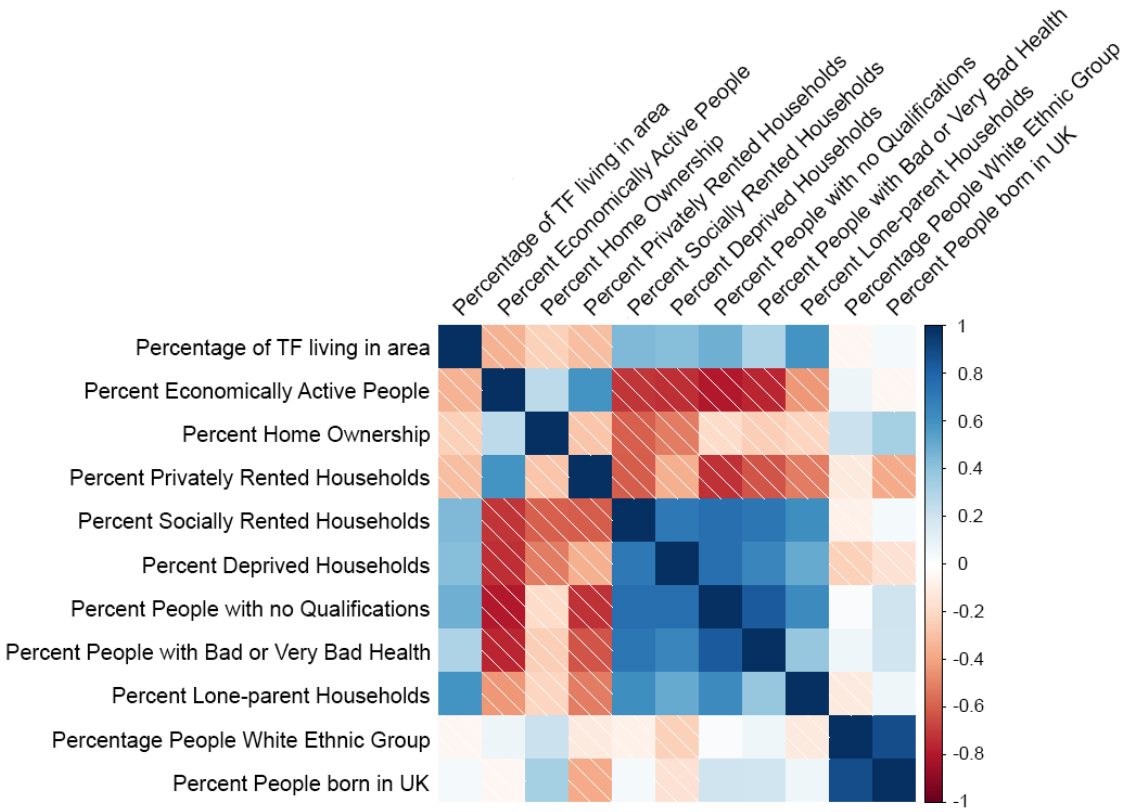


Figure 10: Pearson Correlation between various characteristics of the city's Output Areas (using Census 2011 data) and percentage of TF living in the Output Area (using ECC data)

The percentage of TF living in an area was negatively correlated with an area having economically active people (-0.35), and positively correlated with an area having people with no qualifications (0.48). This suggests that greater proportions of TF live in areas with lower levels of economic activity (employment, etc.) and areas that have higher levels of people with no qualifications. Looking more closely at the complete economic activity data (not plotted), it would appear that greater proportions of TF also live in areas where higher levels of people are employed part-time (0.39), rather than full-time (-0.35) and also where higher proportions of people stay at home to look after their families (0.55). The Census data for economic activity was compiled by asking the household reference person about their economic activity in the week prior to the 2011 Census. A

person (aged 16 or over) was described as economically active if they were in employment, actively looking for employment or about to start employment.

It is likely that higher proportions of TF live in areas with more social housing, less home ownership and fewer households who rent privately; this is evidenced by the positive correlation with the percentage of households living in social housing (0.45) and the negative correlations with the percentage of households who own their own home (-0.23) and households renting privately (-0.31).

Higher proportions of TF would appear to live in areas with higher levels of household deprivation and bad general health; this was implied by the positive correlations with the percentage of households who were deprived in at least one dimension (0.43) in an area and with the percentage of people with bad or very bad general health (0.30) in an area.

There were four household characteristics that were considered to be indicators of household deprivation by the 2011 Census (Office for National Statistics, 2014), these were:

- Employment: any member of the household (who is not a full-time student) is unemployed or long-term sick
- Education: no person in the household has qualifications greater than Level 1, and no person aged 16-18 is a full-time student
- Health and disability: any person in the household has general health classed as 'bad' or 'very bad', or has a long-term health problem
- Housing: the household has overcrowded accommodation, or is a shared dwelling or has no central heating

The attribute that was most highly correlated (in Figure 10) with the percentage of TF living in an area was the percentage of lone-parent households in an area; the correlation was 0.59. This would imply that TF tend to live in areas with higher proportions of lone-parent households. One other notable aspect was the lack of correlation with being born in the UK (0.04) or belonging to the white ethnic group (-0.05) which would indicate there was little identifiable pattern here. The highest correlation the percentage of TF living in an area had with any ethnic group was for the percentage of black/African/Caribbean/black British people living in an area (0.29). Similarly, the highest correlation with any place of birth was for the percentage of people living in an area who were born in Africa (0.30).

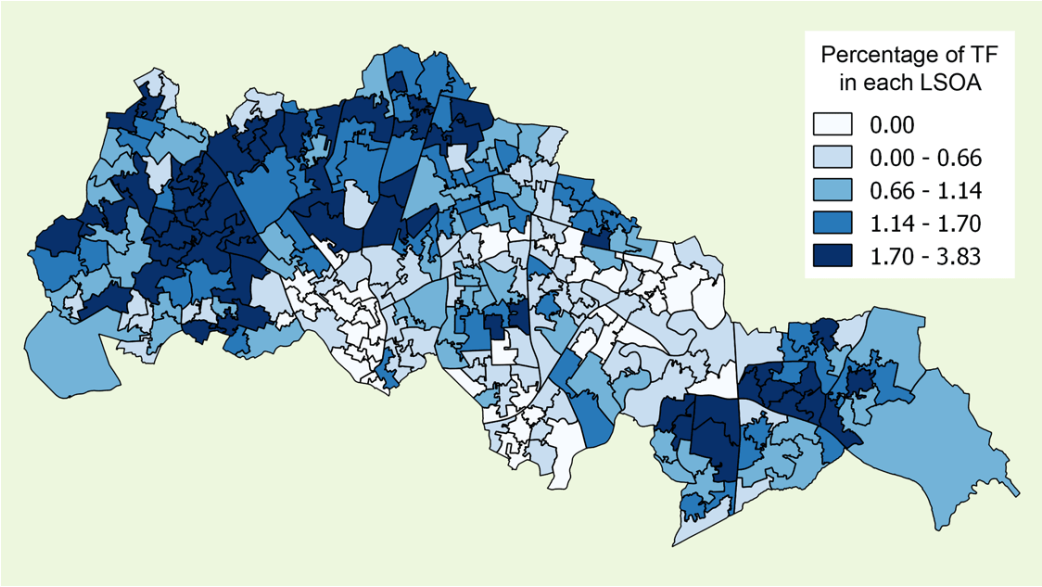
Whilst the correlations provide useful insight into the types of areas that the TF lived in at the start of their intervention treatment, these statistics could not provide any information as to the proximity or geographical representation of the families, therefore Figures 11 to 19 plot some of these attributes on maps to visualise and compare the various demographic statistics with TF location.

Plots utilising the LSOA are shown as not all the data was available to the OA level; and the less fine-grained nature of these plots (compared to the OA level) was easier to analyse visually. No identifying geography is included in the maps in order to protect the anonymity of the city. Figures 11 to 16 utilise the 2011 Census data, and Figures 17 to 19 utilise the Police crimes data and ECC crimes data (taken for the year 2011 to match the Census data). On each page, a plot of the percentage of TF in each LSOA is also included (as in Figure 8) so that the concentration of TF in an area might be easily compared to the particular characteristic.

Figure 11 plots the percentage of households in each LSOA that were classed as deprived in at least one dimension by the 2011 Census. When compared to the location of TF, it appears that higher proportions of TF live in the areas that contain proportionally more deprived families, as was also indicated by the correlation. There was a similar pattern for the proportion of people with no qualifications (Figure 12); it appears that, in general, higher proportions of TF live in areas that have higher proportions of people with no qualifications.

Where considering general health (Figure 13), LSOAs with a higher proportion of people who considered their health to be 'bad' or 'very bad', also tended to align with those LSOAs that contained proportionally more TF. This was true also of lone parent households; of all the plots, Figure 14 and the map detailing the percentage of TF in each LSOA appear to be the most similar, confirming the high correlation between these two attributes. Conversely, Figure 15 appears to confirm the negative relationship between the proportion of TF living in an area and home ownership.

Figure 16 plots the percentage of households in each LSOA whose household reference person was considered economically active in the week before the 2011 Census. It was notable that greater proportions of TF tended to be located in areas with less economic activity (as also indicated by the correlation); this was particularly evident where the North-Western block of TF were considered.



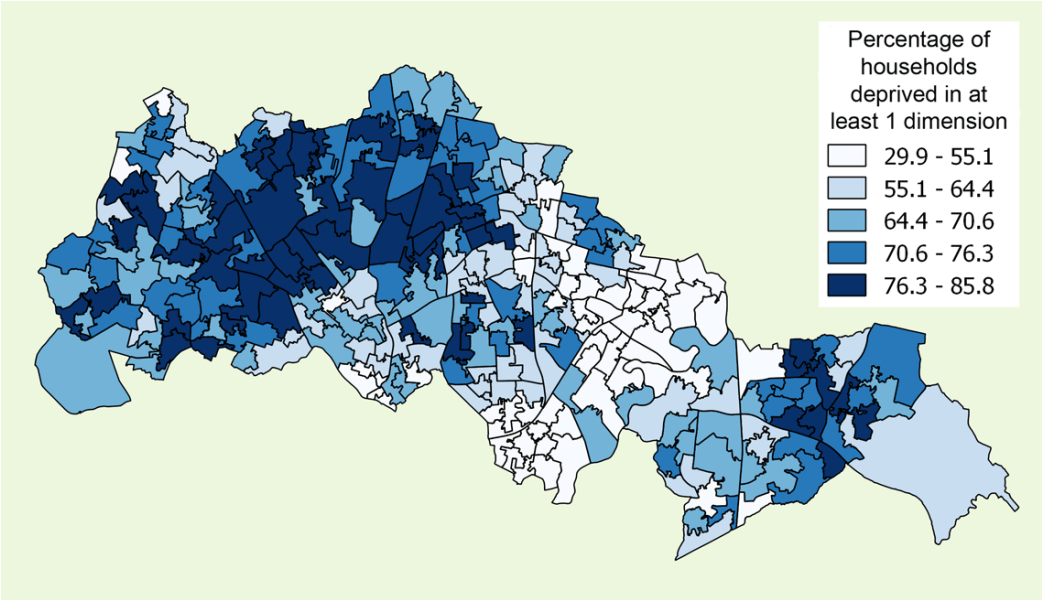


Figure 11: Percentage of deprived households per LSOA (Census 2011 data)

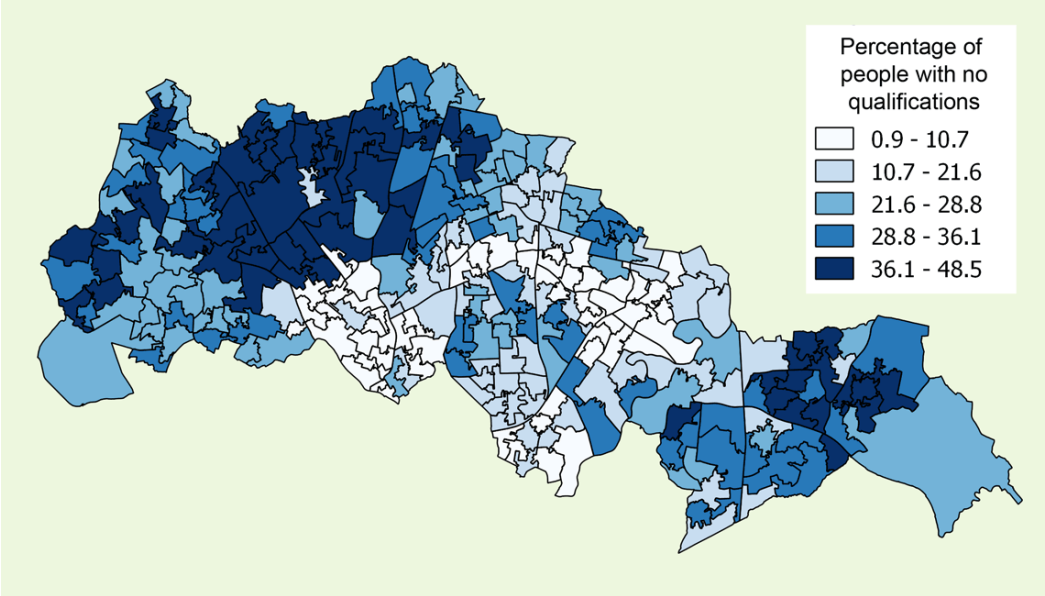
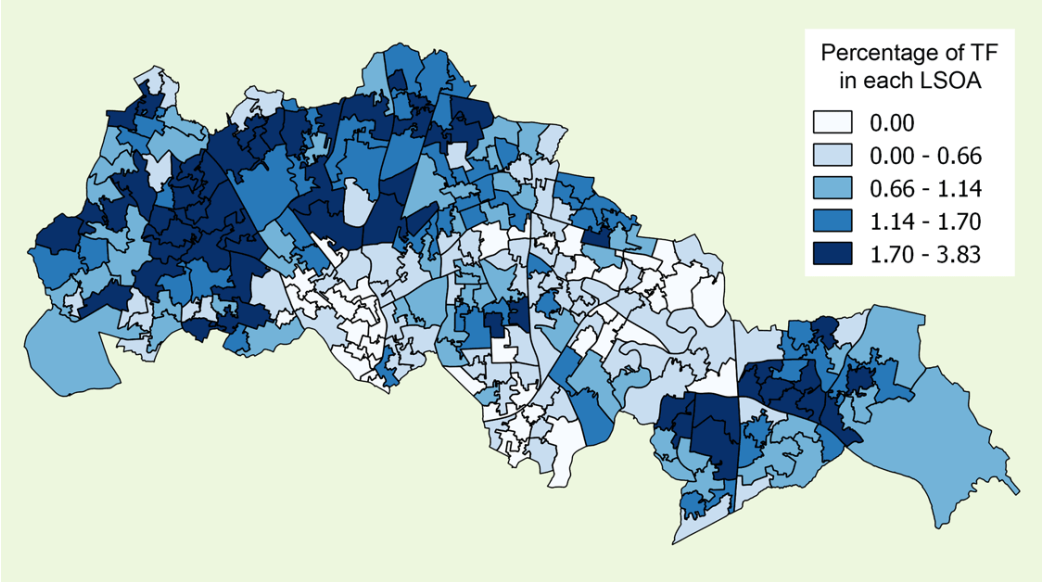


Figure 12: Percentage of people with no qualifications per LSOA (Census 2011 data)



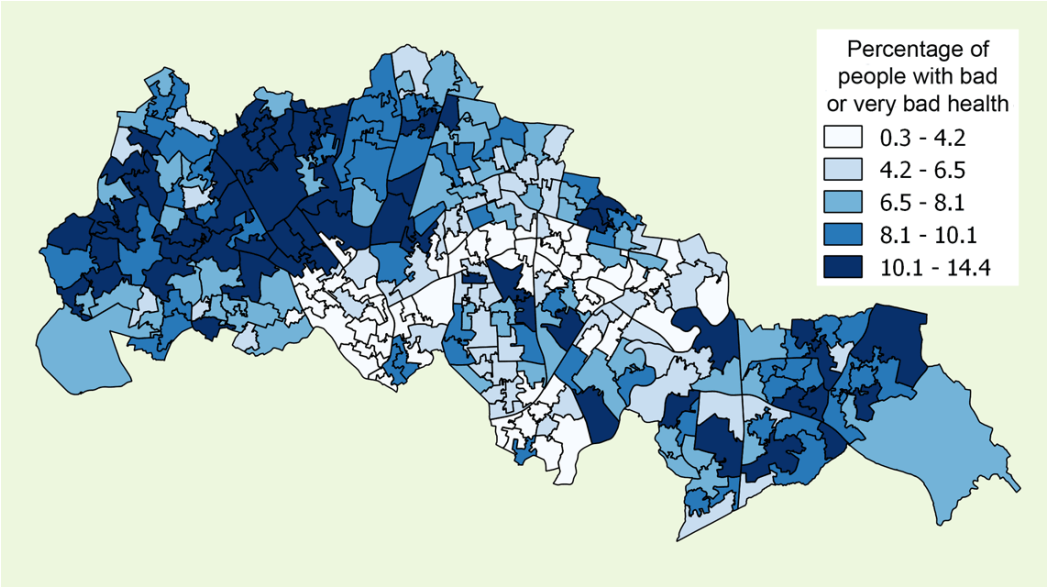


Figure 13: Percentage of people with bad or very bad general health per LSOA (Census 2011 data)

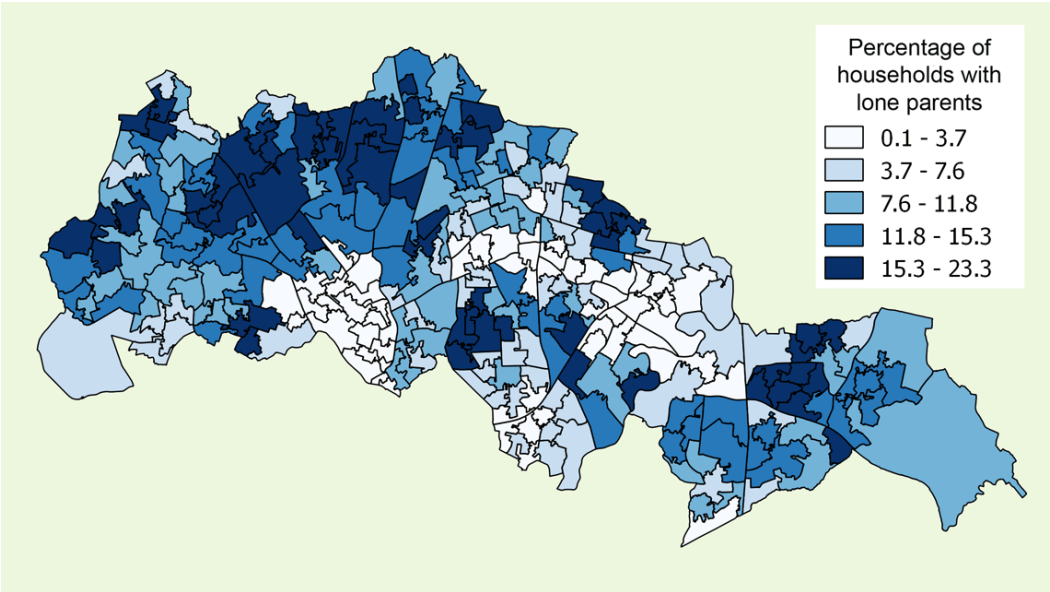
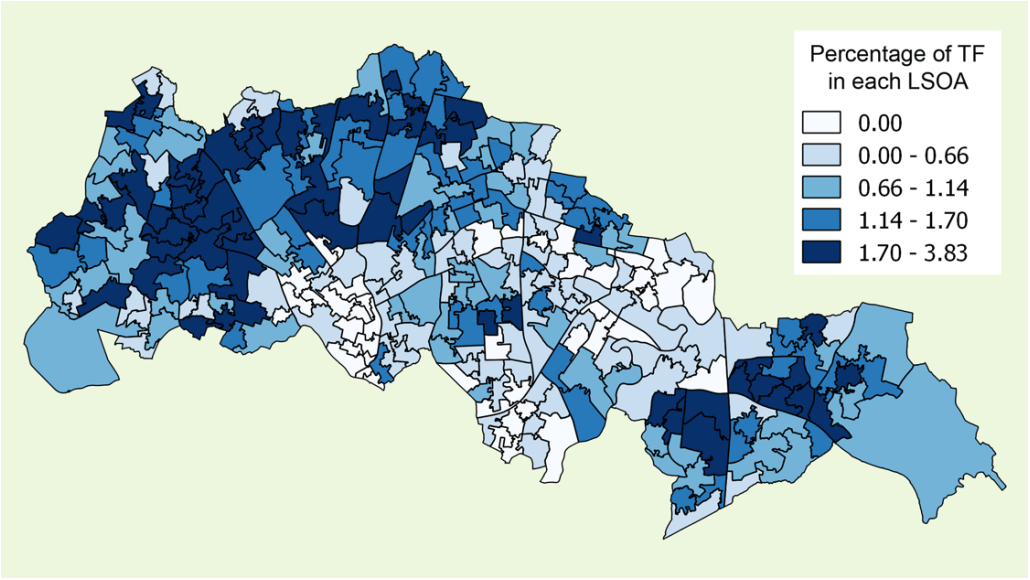


Figure 14: Percentage of lone parent households per LSOA (Census 2011 data)



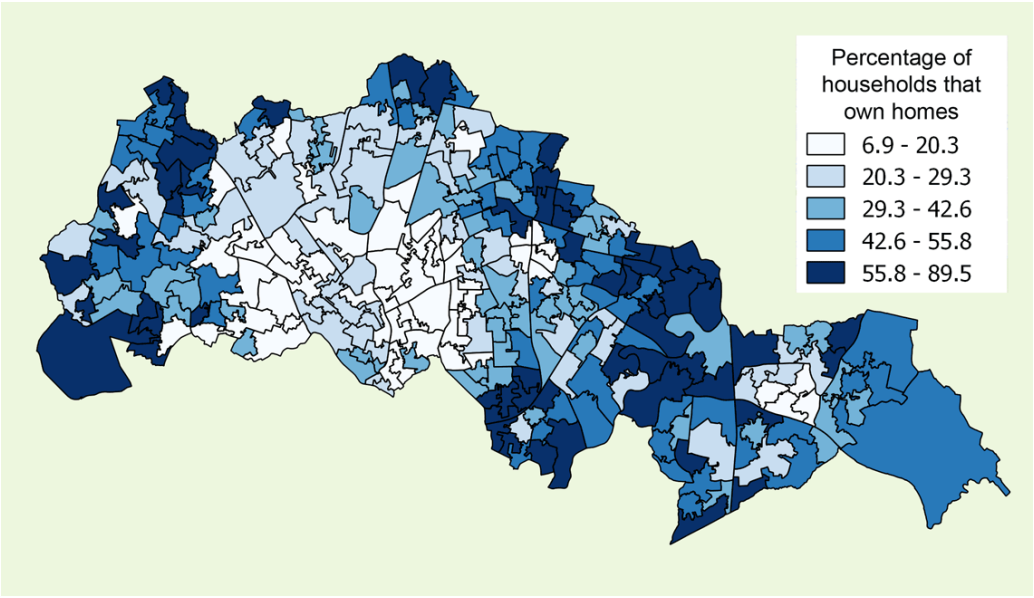


Figure 15: Percentage of households that own their home per LSOA (Census 2011 data)

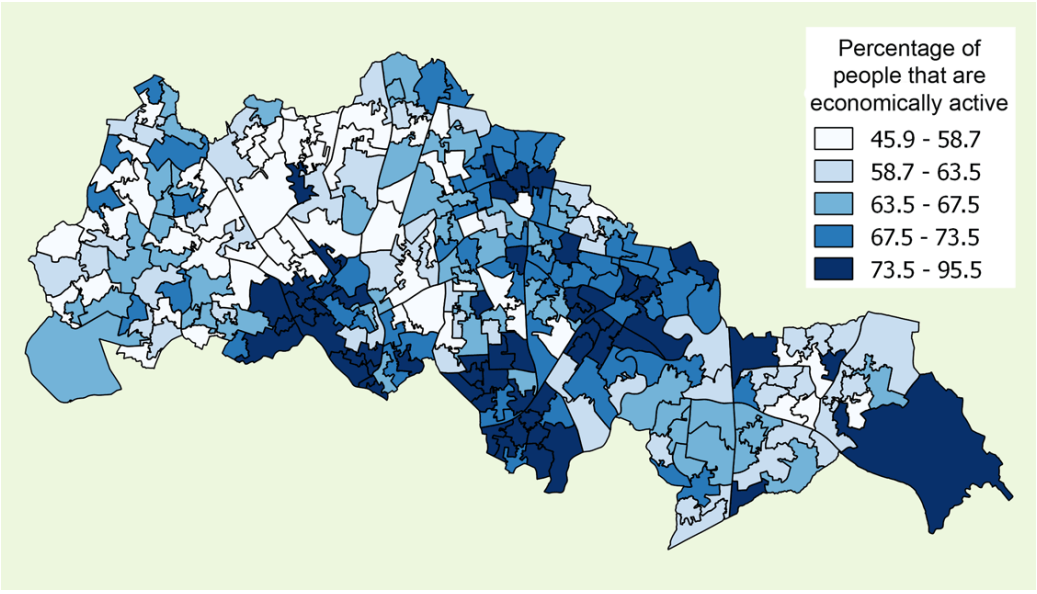


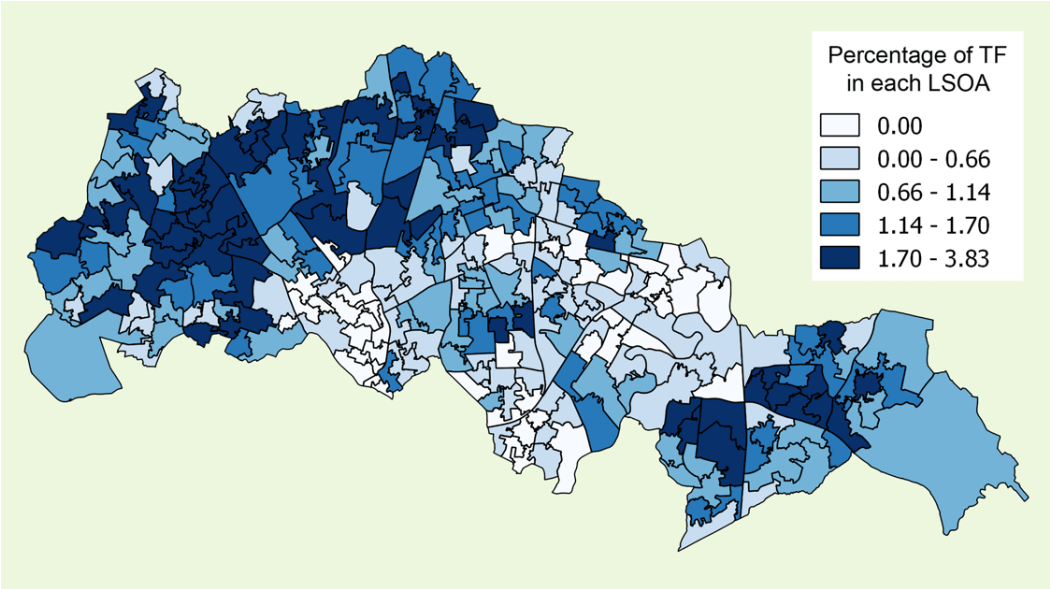
Figure 16: Percentage of economically active people per LSOA (Census 2011 data)

In terms of crime, Figure 17 plots the percentage of people living in each LSOA who committed a crime in 2011. This was derived from the ECC crime data; each crime was linked to one or more persons, which could then be linked to that person's address. However, not all crimes could be linked to a person's address, therefore this only considers the data that could be linked. To derive the percentage, the total number of individuals living in each LSOA who had committed a crime in 2011 was divided by the total number of people living in that LSOA (using the census 2011 count). The correlation between the percentage of TF living in each LSOA with the percentage of people living in each LSOA who had committed a crime was 0.57, indicating that TF tend to live in areas that contain higher proportions of people who have committed crimes.

In contrast to this, Figure 18 plots the percentage of total crime occurring in each LSOA, utilising the police data (Home Office, 2016). There was no correlation (0) with the percentage of TF living in an area, indicating that there was no detectable relationship (in terms of correlation) between where TF live and the amount of crime committed in the area, and this was also indicated by the plot. Figures 17 and 18 (and the correlations) suggest that TF live in areas that have higher proportions of people who have committed crimes, but that there is no correlational indication of any pattern with respect to the amount of crime committed in these areas.

Overall, the correlations and the visualisations of the various characteristics compared to where the families lived indicate that at the start of their first intervention TF tended to live in areas with higher percentages of lone-parents, higher levels of deprivation, lower educational levels, poor health, less economic activity and higher levels of social housing. To some degree, such results might be expected, but it was useful to consider these characteristics in relation to the geographical proximity of families and areas.

Whilst this 'place-based' analysis has already been informative, it was performed in order to build the groundwork for later analysis which considered whether the particular cluster assignment had any relationship to where a family lived. When analysing a large group of families (or any large group of data) there is always the possibility that there could be an averaging effect; identifying different clusters may minimise this effect and allows the possibility of discovering 'place-based' characteristics that are specific to particular clusters, and therefore potentially providing further insight into the families.



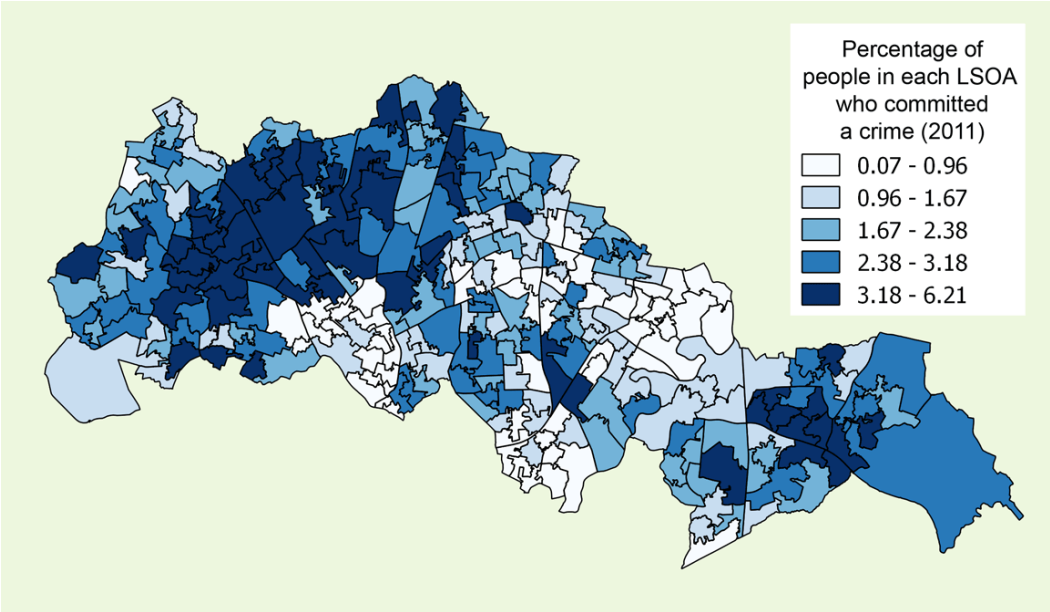


Figure 17: Percentage of people who committed a crime in each LSOA (2011), using ECC data

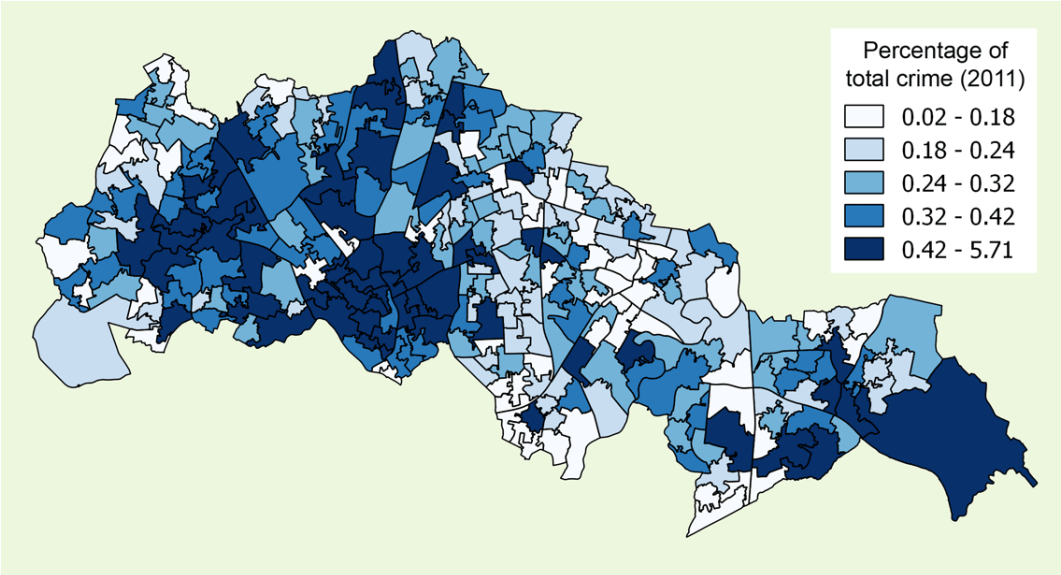


Figure 18: Percentage of total crime occurring in each LSOA (2011), using Police data (Home Office (2016))

6.2.4 Hierarchical Clustering Preparation

The data was clustered in order to identify unique groups of TF. The data for the year leading up to a family's first intervention was utilised in order to particularly consider what precipitated a first intervention. Therefore, nothing pertaining to the outcome of treatment was included, only that which was known about the family prior to the start of treatment. It was thought the identification of these groups might lead to a greater understanding of the types of different families that exist within the data, and therefore a better understanding of their needs. In general, discovery of unique groups in the data might allow for more detailed analyses of the particular groups, rather than the overall global analysis of the families which may succumb to an averaging effect.

This was a data-driven analysis, and the aim was to include as much data that described the family as was possible in the cluster analysis, without having to make prior assumptions. After considering the data and the various shortcomings of some of the attributes (incompleteness, etc.), the attributes remaining for clustering were:

- School Absence: Percentage of unauthorised school absence overall for applicable members of the family
- School Exclusion: count of school exclusions for the whole family
- Children in Need events (CIN): binary
- Child Protection Plan event (CPP): binary
- Looked After Child events (LAC): binary
- Criminal Offences committed by adults (aged 18 or over on the first intervention date): count
- Criminal Offences committed by children (aged under 18 on the first intervention date): count
- Members not in education, employment or training (NEET): binary

The NEET, CIN, CPP and LAC attributes were included as a binary indication of whether a family had any of these events in the year prior to intervention as a count of these events did not make logical sense. For instance, if using a count of NEET events, it would have in most cases simply indicated the same individual moving in and out of NEET status. The same was true of the CIN, CPP and LAC data; many of the events were inter-related and a count did not make sense.

The percentage of school absence could be calculated by counting the total number of unauthorised school sessions divided by the total number of available sessions for the whole family. An unauthorised session is one that a child did not attend, and the school did not authorise this absence (absence might be authorised if a child was ill, for example). This was only performed where school absence data was available; if only one child attended school, then only their data was compiled, if there were two or more children their data was combined and compiled. School exclusions and criminal offences could logically be included as a count. The Children Missing Education (CME), Anti-Social Behaviour Legal Actions (ASB), inclusion in a Pupil Referral Unit (PRU) and social care Drug/Alcohol data were excluded from further analysis as they were incomplete and very sparse. Benefits data was also excluded, as it was not complete and could not be accurately attached to the time period before a family's first intervention.

The eight attributes that remained for clustering meant that the focus of the clusters would be around child safeguarding (CIN, CPP, LAC), education (school absence, exclusion and NEET status) and crime (committed by adults and children).

In order to visualise the overall trend of the types of events that each family had in the year prior to intervention, a heatmap was plotted, Figure 19. Plotting the data in this way can be useful as it may indicate groups within the data. The plot provides a binary indication of each of the events, i.e. simply whether a family had that event or not. Each of the 2155 families was represented as a very thin vertical line, running from the top of the plot to the bottom. Purple indicates the presence of an event, whereas turquoise indicates the absence of an event. Highlighted on the right of the plot, the block of turquoise indicates that 605 families had none of the listed events occur in the year prior to their first intervention. There were also three other distinct blocks of families, as highlighted: those with only CIN events (243); those with only school absence (223); and those with only CIN events and school absence, but nothing else (182). These groups of families could be thought to form their own clusters as they represent distinct groupings, and as such were excluded from the cluster analysis. The criteria used for deciding which groups to exclude was that a group must be of a reasonable size; greater than 10% of the data (where the zero group had already been removed), i.e. it must contain more than 155 TF.

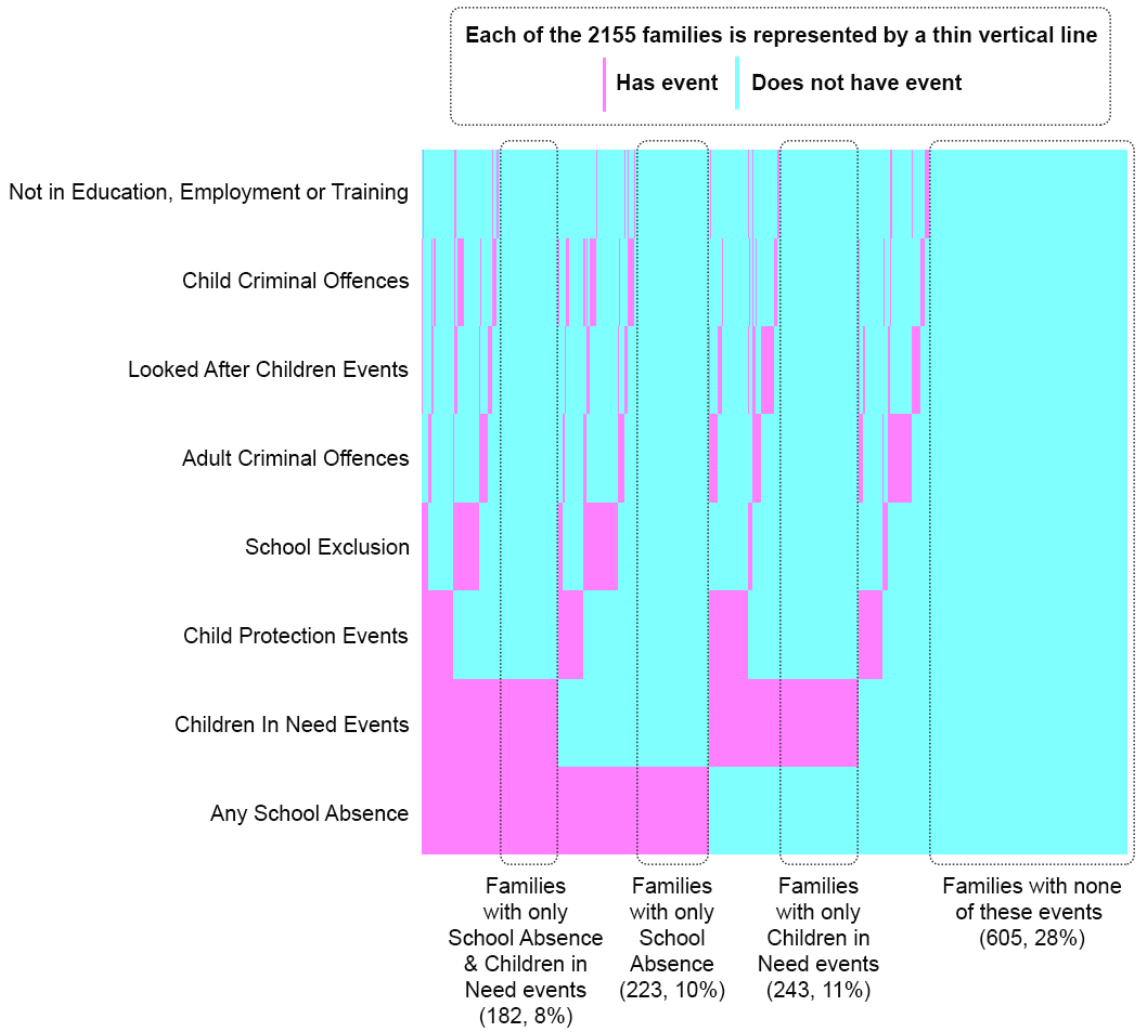


Figure 19: Heatmap of the events occurring for each TF in the year prior to intervention, using ECC data

As Figure 19 highlights, 28% of families had none of the specified events linked to them; 605 out of 2155 TF did not have a record of any of the eight events occurring in the year prior to their first intervention date. However, had the DWP benefits data been reliable and included as an event, it would have accounted for 40% of these families. For comparison, 44% of families who had at least one event also received DWP benefits, so the proportions for both groups were somewhat similar. The other attributes that were excluded due to data unreliability (PRU, CME and ASB) only accounted for 0.2% of the families with no events.

Of the 605 families with no events, it was noted that 20 families (3%) had an existing Child Protection Plan (issued over a year before the first intervention date), and that 6 families (1%) had Looked After Children events (occurring over a year before the first intervention date and that appeared to be ongoing). However, since the focus was on events occurring in the year prior to intervention (and what might have precipitated a family's

need for an intervention), these families were left in the group with no events, as they had not had any documented events in the previous year; that is, their status remained the same.

Overall, Figure 19 highlights that while there were some families with a complex assortment of problems and events, there were others with only a single problem, or none at all, at least where considering the available data. Whilst this appears to contradict the Government's criteria that TF have multiple problems, it must be considered that the available data does not cover all the criteria. No family had all eight of the different events occur in the year prior to intervention, and only nineteen families (0.8%) had 5 or more of the different events. The majority of families (1756, or 81%) had two or fewer different events occur. Table 9 details the percentage of families with each number of different events. It is also worth considering that in Figure 19 any amount of school absence was considered as an event; this could mean that a family was marked as having school absence when they might only have one child who had one unauthorised school session.

Table 9: Percentage of families with each number of different types of events in the year prior to intervention (ECC data)

	Number of different types of event a family had								
	0	1	2	3	4	5	6	7	8
Percentage of families (number in parentheses, total 2155)	28.1% (605)	29.7% (639)	23.8% (512)	13.2% (285)	4.4% (95)	0.8% (17)	0.1% (2)	0	0

Considering the Government's guidelines as to qualification for the TF programme (a family must have three of the following events: have members involved in crime or anti-social behaviour; have children not in school; have an adult on out of work benefits; or cause high costs to the public purse), it would appear that some of these families may not meet the criteria. However, since not all data pertaining to these events was available for analysis, it may be that at least some of the families with no events could have satisfied the criteria were that data available. In particular, the fact that there was no anti-social behaviour data and incomplete benefits data means that one or two of the criteria could have been invisibly satisfied, however the available data cannot reflect this. It should also be considered that the 'local discretion' (or high cost to the public purse) criterion could also have covered events not contained in the data (such as health problems).

One other issue to consider, in terms of the lack of events counted for families, is that some families move in and out of areas, and it is unclear from the available data when this happens. For instance, if a family had lived in another area previously there would be no data (i.e. no school records, etc.) for them contained in the ECC database; this would be contained in their previous Council's database. And since there was nothing in the available data to indicate the date of a family's arrival in the area, it was impossible to determine whether they simply had no events or had none recorded because they were not living in the area. ECC indicate that families moving into the area would likely have been referred into the TF programme by their previous area's social services teams and so would have legitimate issues, but their historical events would not be contained in the ECC database.

Without this data, it was impossible to determine how many families with no, or few, events might have had more events occur in the year prior to intervention and so ultimately satisfy the TF criteria. However, this case study could only analyse the available data, and in order to determine if there were any specific characteristics between the groups of families with and without events occurring in the year prior to intervention, a comparison was carried out; Table 10 compares the two groups.

Table 10: Comparison of TF with and without events in the year prior to first intervention, using ECC data

	Families with events (n = 1550)	**Families without events (n = 605)**
Family Size	Range: 1-11, mean: 3.7	Range: 1-9, mean: 2.2
Family Composition	4.6% were single person families 3.7% have no children in family 8.6% have no adults	41.2% were single person families 49.8% have no children in family 8.1% have no adults
Intervention type	AO: 20.3% CFPT: 29.7% FF: 16.4% FINIS: 3.7% FIP: 29.8% Other: 0.1%	AO: 35.0% CFPT: 28.9% FF: 12.2% FINIS: 1.3% FIP: 22.5% Other: 0.0%
Intervention status	Open: 5.4% Planned Ending: 74.7% Unplanned Ending: 19.9%	Open: 5.1% Planned Ending: 74.9% Unplanned Ending: 20.0%
Receiving DWP benefits on first intervention date	43.7%	40.2%
At least one change of address in previous year	48.5%	36.0%

The most notable difference between the two groups was the difference in family size: those who had events occur in the year prior to intervention generally comprised larger families than those without events. Of the families without events, 41.2% comprised a

single person, compared to only 4.6% for those with events. Figure 20 and Figure 21 plot the distribution of family sizes for the two groups and illustrate the difference in distributions: families without events (Figure 21) are right-skewed with proportionally more families having only one member; families who had events (Figure 20) comprise a more Normal like distribution with proportionally more families having 2, 3 or 4 members.

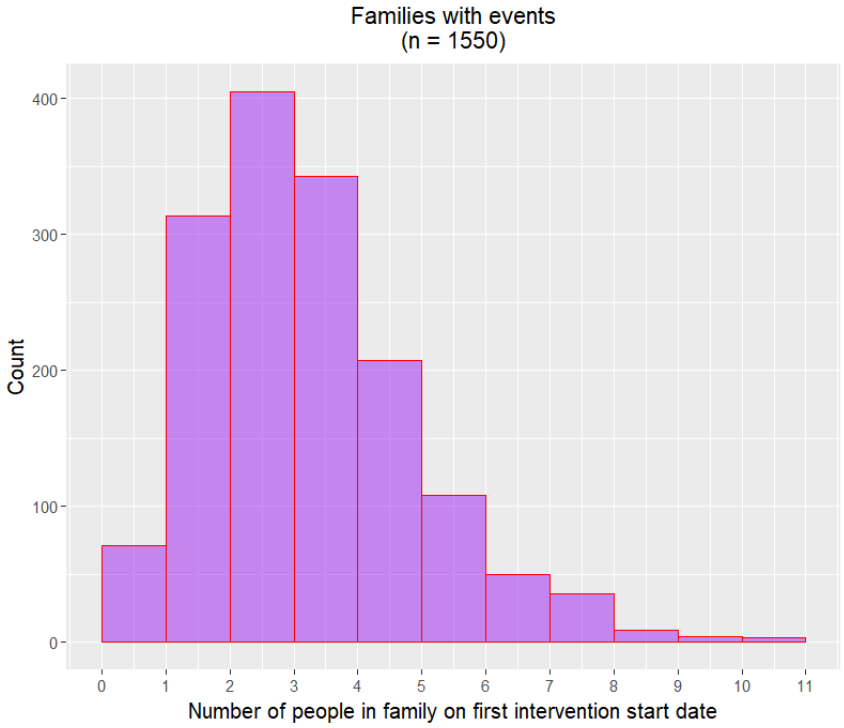


Figure 20: Distribution of family size for TF who had events in the year prior to intervention, using ECC data

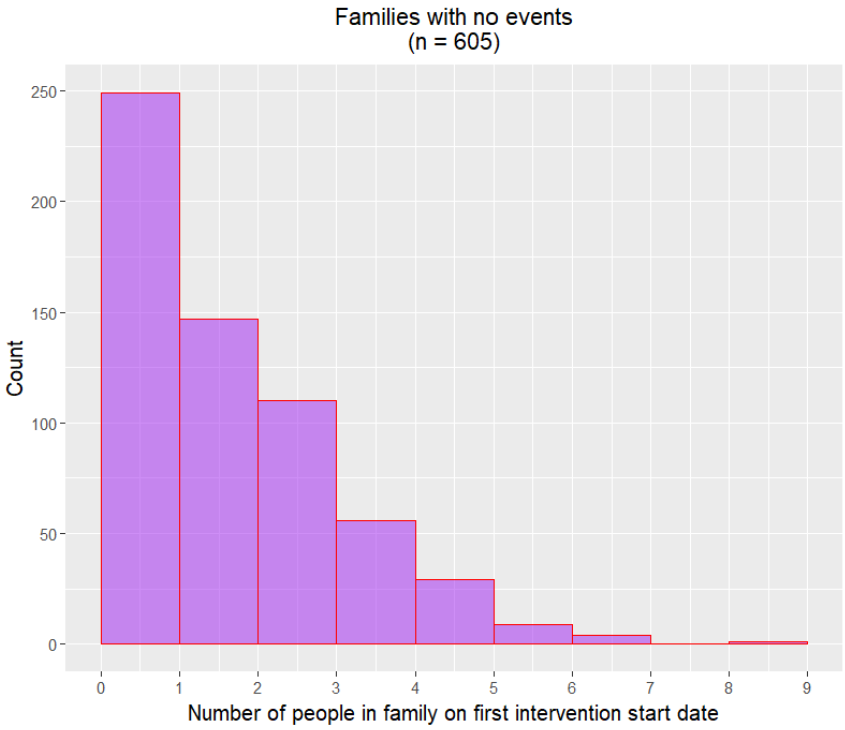


Figure 21: Distribution of family size for TF who had no events in the year prior to intervention, using ECC data

Something that may at least partially explain the number of families with only one member could be problems with the data itself. In general, TF should not comprise one-person families, and it may be that due to missing linking attributes some individuals simply were not linked to the rest of their family. The ECC acknowledged that there were problems such as this within the data, although it was not clear how pervasive they were.

Another notable difference between the two groups was that almost half (49.8%) of the families without events had no children (aged under 18), compared to 3.7% for those with events. The fact that there were fewer children in the no event group might help to explain the absence of events; most of the events concern issues with children (school issues, and child safeguarding), and if there were fewer children in the group then it was likely that there would be fewer of these events. The only event that would be considered specifically adult was criminal offences committed by adults.

Where considering Table 10 and the comparison between groups, it was also notable that a higher proportion of families without events received AO treatment (35% compared to 20%). AO treatment is provided to families who are deemed at risk of developing more complex needs, so this may provide some insight into the families (or at least those who were receiving the AO treatment). It may indicate that whatever their particular problems were, they could have been fairly low-level and these families were receiving treatment in order to prevent escalation.

Another notable difference between the two groups was that the families with no events had fewer address changes than those with events (36% of families had at least one change of address in the year prior to intervention compared to 49%). However, it is possible that this difference may have been a reflection of how the data was collected. If families had no events, then there simply may have been less opportunity to record a change of address since there would be less interaction with the services that collect this type of data. But it is also possible that a characteristic of this group could be that they just do not change address as frequently.

Perhaps surprisingly, despite the significant differences between the two groups, the outcomes of the first intervention treatment were almost identical across both groups. The proportion of planned endings was just under three quarters for both groups, and a fifth of interventions had unplanned endings for both groups. This appears to imply that the differences between the groups had little bearing on how an intervention would end.

Whilst this part of the analysis concentrated on the unique group of families with no events, there were three other large groups of families identified within the data that were also considered to form unique groups. These were families who just had school absence (223 families), families who just had CIN events (243 families) and families who just had school absence and CIN events (182 families). Together with the no event group (605 families) they were excluded from the cluster analysis and were considered to form their own unique clusters; they are referred to as the 'pre-specified' clusters throughout the rest of the analysis.

6.2.5 Models

Hierarchical clustering was performed upon the data (minus the pre-specified clusters) in order to identify groups of similar families. This method was chosen as it can effectively deal with mixed data (that is, data that contains both numerical and categorical attributes). It also provides a visual clue as to how the data is formed into clusters and is particularly useful since there is no requirement to know in advance how many clusters the data contains. Gower's general coefficient of similarity (J. C. Gower, 1971) was used to calculate the dissimilarities. This normalises the numerical data by dividing by the range, therefore, to avoid skewing the data, any extreme outliers were capped (at the next lowest value) prior to this. There was no indication that the few outliers identified were due to incorrect data, so it was not necessary to exclude them from the data; it was felt that retaining them at a capped level would still represent that they had a high value.

Initial experiments found that utilising mixed data resulted in the binary attributes being heavily favoured by the hierarchical clustering algorithm, therefore it was necessary to apply weights to the binary data to lessen its impact. The binary attributes were given half the weight of the numerical attributes.

The Complete-Linkage method was chosen as it can produce compact clusters, and is suitable for mixed data. The other hierarchical clustering linkage methods were also utilised in order to determine whether they might produce similar results. Experiments with various attribute combinations and different types of clustering were performed as part of the data analysis stage, in particular a consideration of using all binary data, but it was felt that this unnecessarily removed detail from the data.

Cluster size was decided by visualising the dendrogram of the hierarchical clustering and considering the silhouette values (Rousseeuw, 1987) and Goodman and Kruskal's Gamma

coefficient values (Milligan and Cooper, 1985). Visualisations of the resulting clusters mapped back to the data on a two-dimensional representation were plotted using T-Distributed Stochastic Neighbor Embedding (t-SNE) with the 'Rtsne' R package (Krijthe and van der Maaten, 2017).

Since the preceding data analysis identified large groups of TF that all shared the same characteristics in the data (i.e. families that had no events, or just a single event prior to intervention), these pre-specified groups were excluded from the clustering, and were considered already as pre-formed clusters. These pre-specified clusters accounted for 1253 TF, leaving 902 TF records for the cluster analysis.

In order to provide deeper insight into the clusters, decision tree learning was utilised to derive rules for the cluster assignments. This was utilised with the 'rpart' R package (Therneau et al., 2017), which is an implementation of the CART algorithm.

Analysis was performed to determine if geographical location might be a factor in determining which cluster a TF belongs to. QGIS, which is a geographical information system (GIS) software was utilised for mapping, and machine learning was performed to determine if place-based data was a predictor of cluster membership. Decision tree learning (with the 'rpart' R package), random forests (with the 'randomForest' R package (Liaw et al., 2015)) and generalized boosted models (with the 'gbm' R package (Ridgeway, 2017)) were utilised. To provide a comparison to a more traditional regression model, multinomial logistic regression was also performed (with the 'nnet' R Package (Ripley and Venables, 2016))

6.3 RESULTS

6.3.1 Hierarchical Clustering using Complete-Linkage

Clustering was performed upon 902 records. The Complete-Linkage method produced seven clusters, where cutting at a height of 0.58, which is the point at which the greatest drop in height occurs in the dendrogram (Figure 22).

Upon visual inspection of the dendrogram, it appeared that a cut creating five clusters might also be a viable solution, however, whilst it can be informative to study the dendrogram, it is not a very scientific method of deciding the cluster number. Therefore, in order to more rigorously determine which was the optimal point at which to cut the

dendrogram, the silhouette values and Goodman and Kruskal's Gamma coefficient values were calculated for various cluster solutions (Table 11). These metrics were chosen as they are suitable for mixed data; many are only suitable for data that can be represented in Euclidean space. For both measures, a higher value is desirable. They both can take values from -1 to +1, and a value closer to +1 indicates better cluster cohesion.

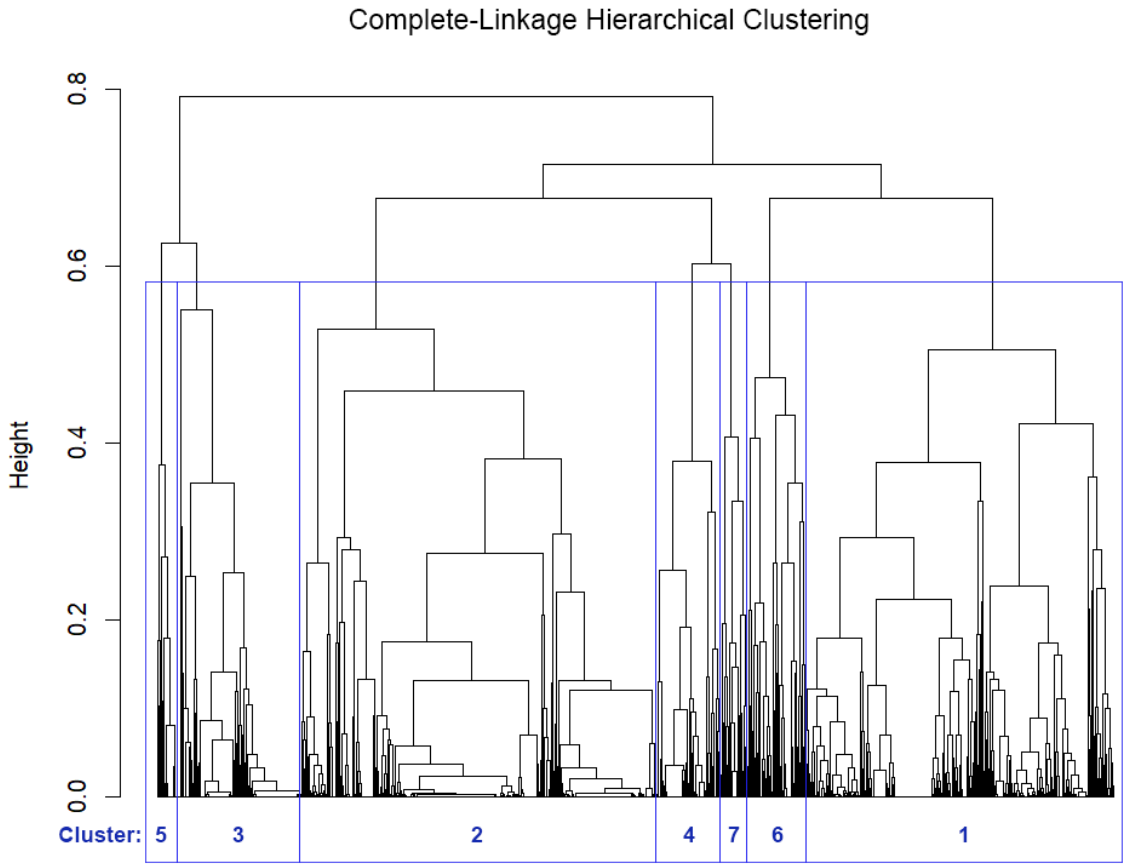


Figure 22: Complete-Linkage hierarchical clustering dendrogram with the seven-cluster solution highlighted

Table 11: Comparison of cluster metrics to determine the optimal number of clusters

Number of clusters	Overall Silhouette Value	Gamma Coefficient
2	0.21	0.26
3	0.27	0.39
4	0.27	0.48
5	0.34	0.66
6	0.34	0.68
7	0.36	0.70
8	0.36	0.70
9	0.31	0.71

Table 11 and Figure 23 highlight that for both metrics, the values kept rising up to the 7-cluster solution, they then both hit a plateau before the silhouette value decreased at 9 clusters and Gamma increased slightly at 9 clusters. The seven-cluster solution was

chosen as it had the highest silhouette value, and it represented a point where the increase of both values slowed.

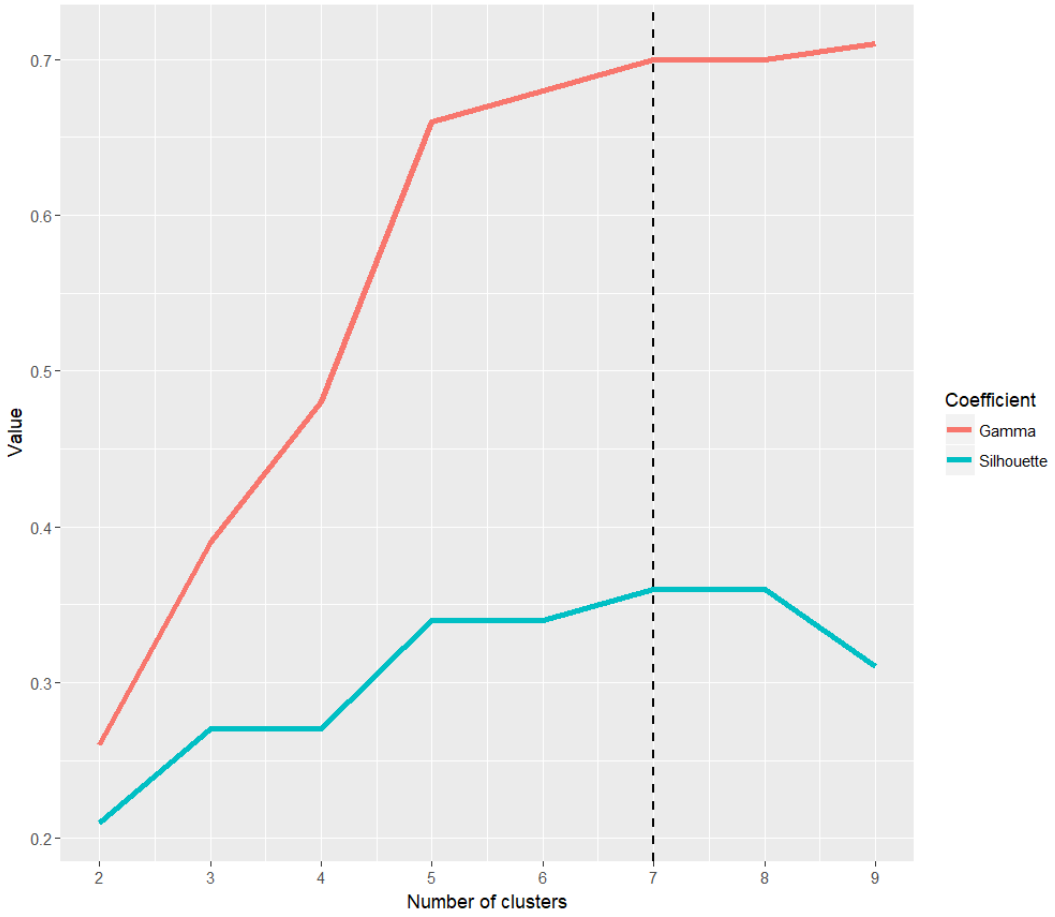


Figure 23: Silhouette values and Gamma statistic values plotted for various cluster solutions using complete-linkage hierarchical clustering method

The other linkage methods were also utilised: Average, Single, Median and Ward's method. For each, a dendrogram was plotted, Figure 24. The Single, Average and Median-linkage methods resulted in dendrograms that were difficult to interpret. Single linkage showed chaining, and the median-linkage method produced inversions, making it impossible to cut the dendrogram in any meaningful way. However, Ward's method showed evidence of more distinct clusters.

The clusters produced by Ward's method (seven clusters) were broadly similar to those found by the Complete-Linkage method. The Adjusted Rand value for comparing the Complete-Linkage solution to Ward's method with seven clusters was 0.65. The Adjusted Rand metric may take a value between -1 and +1, with -1 indicating no similarity at all, and +1 indicating an identical clustering, therefore there was a good level of similarity between the two sets of clusters. However, whilst Ward's method found similar clusters, and to some degree helps to confirm the clusters discovered with the complete linkage

method, it was not utilised in the final analysis. This is because, even though there are documented cases of its usage in this way (Finch, 2005), it does not appear to make logical sense to use a sum of squares method with binary (or in this case, mixed) data.

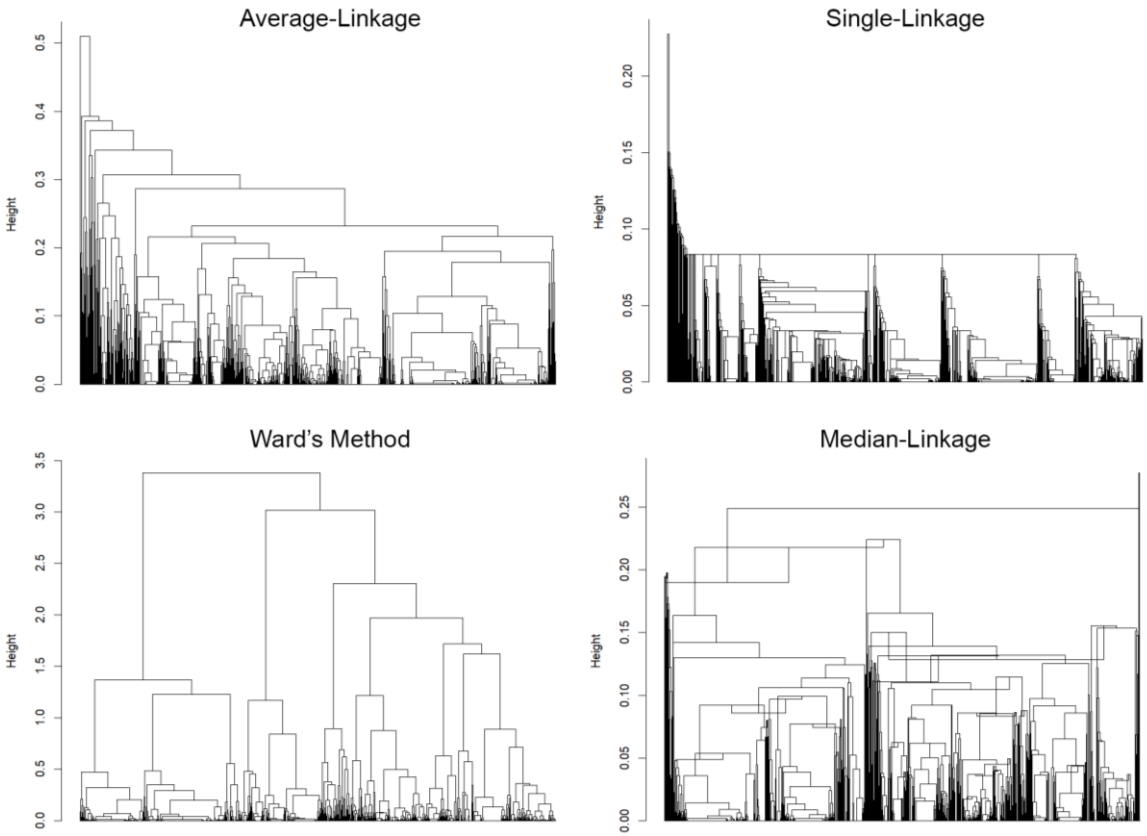


Figure 24: Comparison of other hierarchical clustering linkage methods

6.3.1.1 Details of the seven-cluster solution

As a measure of the clustering quality, the silhouette values for each of the clusters were considered. Values may range from -1 to +1, and are calculated for each object within the cluster. A high value indicates that an object is well matched to its cluster, whereas a low value indicates that it may be better suited to another cluster. The individual silhouette values are averaged for each cluster to produce an overall value for the cluster, as shown in Table 12. The overall average silhouette value for the whole clustering was 0.34.

Table 12: Silhouette widths for seven-cluster solution of the ECC TF data clustering

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Cluster Size	291	335	115	61	21	54	25
Average Silhouette	0.22	0.44	0.56	0.44	0.46	0.14	0.20

The silhouette plot, Figure 25, highlights that cluster 3 was the most cohesive, followed by clusters 2, 4 and 5. The tails on the negative side of the plot represent those records that did not fit as well into their respective clusters. All except cluster 5 had some records

on the negative side, however, clusters 1, 6 and 7 each had a higher proportion (> 10%) of records with negative silhouette values. This is represented by the lower average silhouette values for these clusters and indicates that some caution should be applied in the analysis of these clusters, since they are less cohesive.

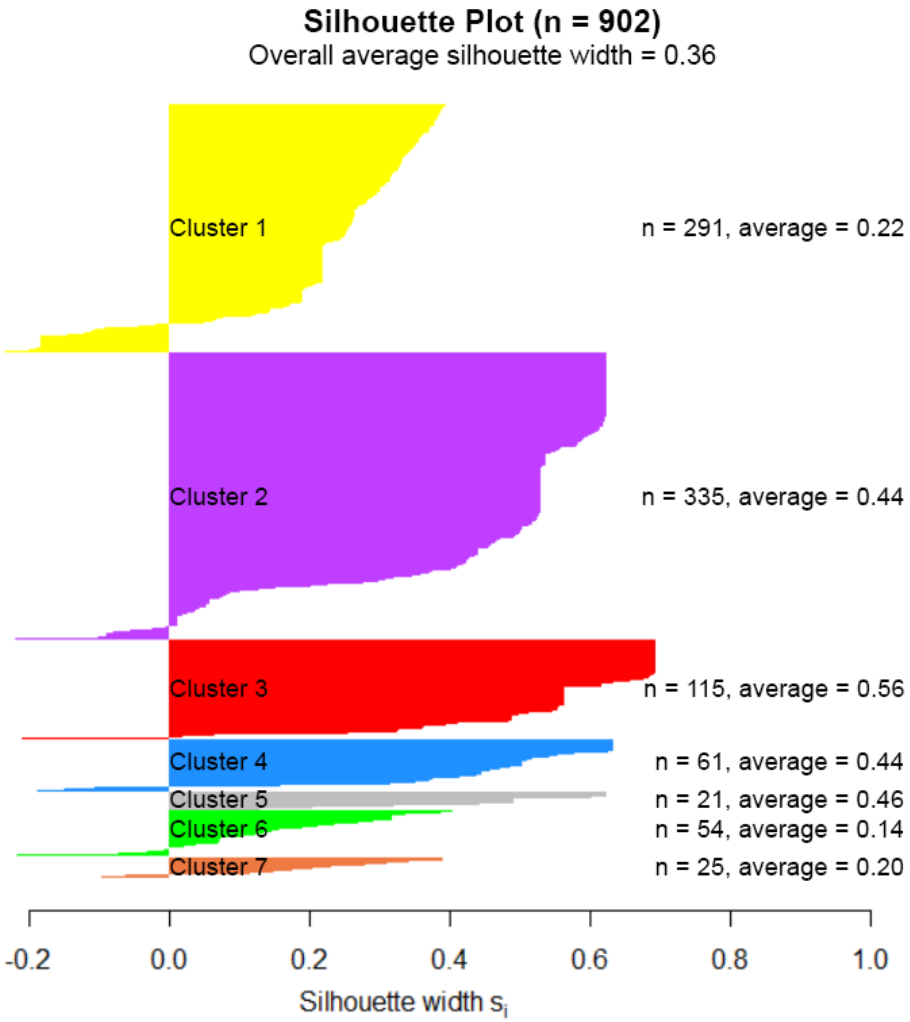


Figure 25: Silhouette plot of the seven-cluster solution, obtained using complete-linkage hierarchical clustering of the ECC TF data

In order to visualise how the clusters relate to the data, Figure 26 plots the clusters represented in two-dimensional space using T-Distributed Stochastic Neighbor Embedding (t-SNE). This was achieved with the parameters set at perplexity (which is approximately equivalent to the number of nearest neighbours) equal to 20, theta (which controls the speed) equal to 0.1, the learning rate set at 200 and using 1000 iterations. This process embeds the eight-dimensions (clustering attributes) of data down into two dimensions and plots it in a scatter plot. Each family is represented by a coloured dot. Although four of the clustering attributes were binary, there was also a numerical count

for each of these attributes contained in the data (e.g. the number of CIN events as opposed to simply having CIN events), and this was used in the plot.

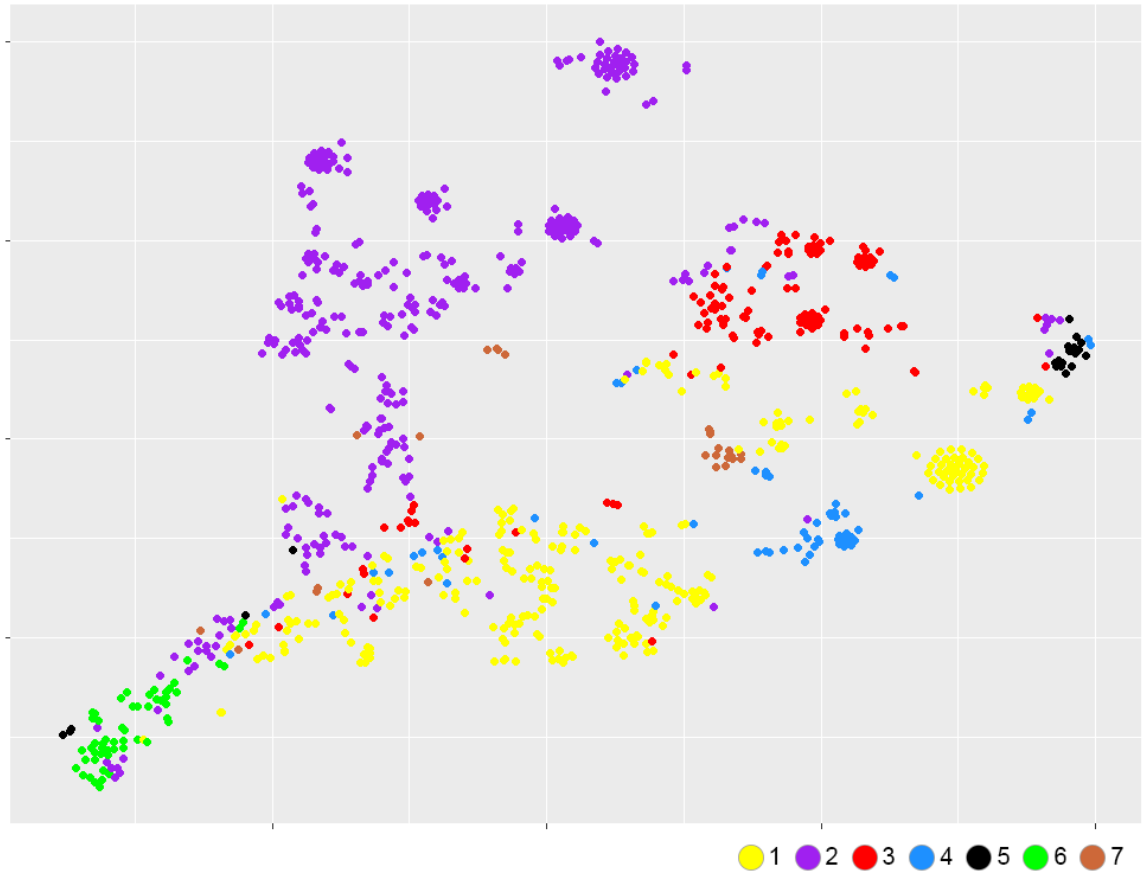


Figure 26: Two-dimensional representation, plotted using t-SNE, of the seven hierarchical clusters obtained from complete-linkage hierarchical clustering of the ECC TF data

Whilst the points from each cluster are not compact in the plot, they do loosely form an overall pattern of togetherness, albeit with some overlaps. Cluster 1 had a low silhouette value (and a larger proportion of families with negative silhouette values) and this appears to be represented by the large spread and overlap of the yellow points in the plot. Interestingly, the points from cluster 6, which had the lowest silhouette value and would therefore be considered the least cohesive group, form a fairly compact group in the plot.

It should be noted that setting different parameters and/or a different random initialisation seed for the t-SNE process will result in a different projection into two-dimensional space, so the method should be utilised with caution, however this method can provide a useful visualisation of high-dimensional data and the results do indicate that there is an underlying pattern in the data with regard to the cluster assignments.

6.3.1.2 Cluster Characteristics

The seven clusters extracted from the hierarchical clustering, together with the four clusters that were extracted prior to clustering (the pre-specified clusters, that were families with: no events; just school absence; just CIN events; and just school absence and CIN events), comprised a total of eleven clusters. Clusters 1 to 7 represent the clusters derived from the hierarchical clustering and clusters 8 to 11 represent the pre-specified clusters. In brief, a summary of their characteristics:

- **Cluster 1: School exclusion and criminal offences**. (n = 291). High levels of school exclusion, criminal offences committed by adults and criminal offences committed by children. Low levels of child safeguarding (CPP, LAC and CIN) events.
- **Cluster 2: Child Protection.** (n = 335). All families had Child Protection events. Very little school exclusion and offences committed by children. Families tended to have younger children with most aged under 11.
- **Cluster 3: Looked after Children.** (n = 115). All families had Looked after Children events, there were low levels of school absence and exclusion
- **Cluster 4: NEET.** (n = 61). All families had members who were NEET, there were low levels of CIN and CPP events
- **Cluster 5: Adult criminal offences.** (n = 21). All families had criminal offences committed by adults, and these were at a high level (with a mean of 4). Almost no child safeguarding (CIN, CPP, LAC) events
- **Cluster 6: High levels of school absence**. (n = 54). All families had school absence, at high levels, with 39% unauthorised absence on average. More complex mixture of events, 78% of families had 3 or more different types of events
- **Cluster 7: Child criminal offences**. (n = 25). All families had criminal offences committed by children, and these were at a high level (mean of 4). Just under half had school absence, and most of those with absence also had child protection plans
- **Cluster 8: School absence only**. (n = 223). All families had school absence but no other events. The average unauthorised absence per family was 6.4%
- **Cluster 9: Children In Need events only**. (n = 243). All families had CIN events but no other events

- **Cluster 10: Absence and CIN.** (n = 182). All families had school absence and CIN events but nothing else. Average unauthorised absence was 10.6%, a little higher than for most other clusters
- **Cluster 11: No events**. (n = 605). All families had none of the events. 41% of families consisted of single people, a far higher percentage than any other cluster. Half of the families had no children

For comparison, a T-Distributed Stochastic Neighbor Embedding (t-SNE) plot was also composed for all eleven clusters, Figure 27. This was achieved with the parameters set at perplexity equal to 20, theta equal to 0.5, learning rate equal to 200 and with 500 iterations.

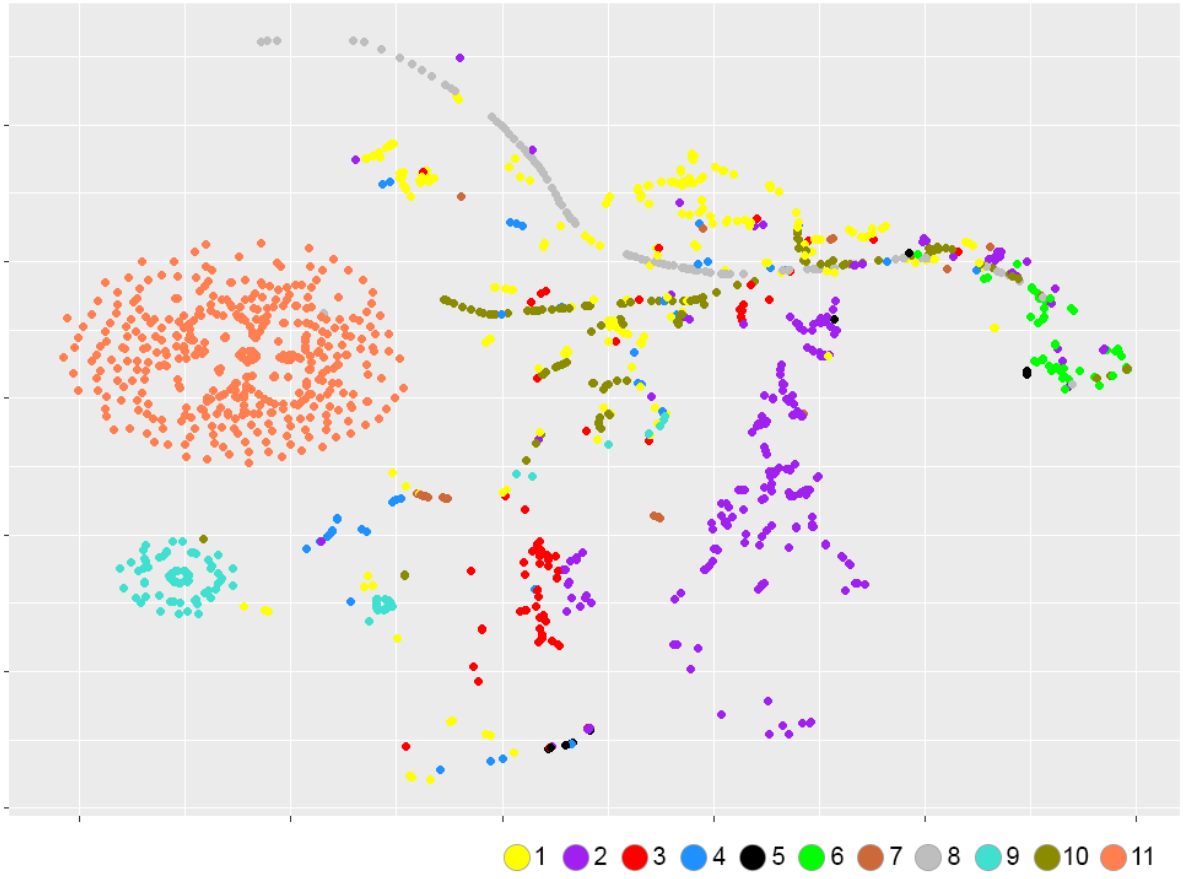


Figure 27: Two-dimensional representation, plotted using t-SNE, of all eleven clusters (the seven clusters obtained by complete-linkage hierarchical clustering together with the four pre-specified clusters)

Although it is difficult to represent 11 clusters in a plot (as it can be difficult to distinguish between 11 different colours) this was attempted in order to illustrate how a t-SNE plot can represent the data and clusters. For instance, cluster 8, which contained only families with school absence, essentially forms a line of dots; this most likely represents the varying levels of absence each family had (some might have 5%, some might have 20%,

etc.). And cluster 11, which was all zeroes (as it was the families with no events at all), is represented by an oval spread of points. Whilst, again, as also illustrated in Figure 26, the clusters do not fall into neat groupings when represented in two-dimensional space using this method, the plot does still indicate underlying patterns in the data with relation to the cluster assignments.

Table 13 details the percentage of families with each event by cluster, with more notable percentages highlighted in bold. For instance, in cluster 2, 100% of families had Child Protection Plans (CPP) in the year prior to intervention. It is clear that the clustering formed a pattern to some degree; aside from school exclusion, each attribute has its own cluster (where every family has that event). For instance, in cluster 5, all families had at least one adult who had committed a criminal offence, and in cluster 4 all families had at least one NEET member, and so on. This may be because many families did not have a diverse mix of events (58% had one or no different events, 81% had two or fewer) and so they fell into clusters that represented their main (or only) issue.

Table 13: Percentage of families with each event per cluster with notable percentages highlighted in bold

Cluster	**1** n=291	**2** n=335	**3** n=115	**4** n=61	**5** n=21	**6** n=54	**7** n=25	**8** n=223	**9** n=243	**10** n=182	**11** n=605
School Absence	66%	44%	30%	43%	29%	**100%**	44%	100%	0	100%	0
School Exclusion	**57%**	6%	11%	16%	24%	54%	24%	0	0	0	0
Children in Need events	42%	58%	63%	31%	10%	57%	32%	0	100%	100%	0
Child Protection events	1%	**100%**	2%	0	5%	22%	44%	0	0	0	0
Looked After Children events	0.3%	12%	**100%**	10%	0	9%	4%	0	0	0	0
NEET	0	2%	1%	**100%**	0	15%	0	0	0	0	0
Adult Offences	**36%**	17%	21%	15%	**100%**	13%	8%	0	0	0	0
Child Offences	**27%**	**4%**	8%	21%	10%	43%	**100%**	0	0	0	0

Table 14: Mean number of events for each cluster, with notable means highlighted in bold

Cluster	**1** n=291	**2** n=335	**3** n=115	**4** n=61	**5** n=21	**6** n=54	**7** n=25	**8** n=223	**9** n=243	**10** n=182	**11** n=605
Absence: Mean percentage of unauthorised absence	3.9	3.7	1.3	2.2	7.8	**38.7**	2.9	6.4	0	10.6	0
Exclusion	**2.1**	0.1	0.2	0.2	0.5	1.1	0.5	0	0	0	0
Adult Offences	0.9	0.4	0.6	0.7	**4.4**	1.5	**4.5**	0	0	0	0
Child Offences	0.4	0.1	0.1	0.3	0.3	1.3	**4.4**	0	0	0	0
Family size	3.7	3.8	3.7	**4.4**	3.3	3.8	3.4	3.6	3.4	3.9	**2.2**

Table 14 details the mean number of events for each cluster, where the count of events was considered. Also included is the mean family size for families in each cluster. Cluster

4 contained the largest families (mean 4.4) whereas cluster 11 contained the smallest families (mean 2.2, as considered previously).

Figure 28 depicts a Nightingale (or Coxcomb) plot of the characteristics of each cluster. Each cluster is depicted by a circle divided into eight equal segments to represent the eight clustering attributes utilised in the model, and each are represented by a different colour. Where a cluster had families with any of the eight attributes, those particular segments are coloured in, with a radius proportional to the value.

For the four attributes that were integer/continuous values (criminal offences committed by children and adults, school exclusion and school absence) the mean value for each cluster is represented. The radius of the segment is proportional to the mean value. For instance, cluster 1 had the highest mean level of school exclusions over all clusters, so this is represented by a fully coloured blue segment; the other clusters that had exclusion had much lower levels and so are represented by a smaller radius segment.

For the four attributes that were binary values (CIN, CPP, LAC and NEET), it was not possible to calculate a mean value, therefore the proportion of families with that attribute/event in each cluster was represented. For instance, in cluster 2 all families had CPPs, and this is represented by a fully coloured orange segment; the smaller percentage of families with CPPs in other clusters is represented by much smaller orange segments.

Using the plot to compare the clusters, the prevalence of CIN events in all but clusters 8 and 11 is perhaps most notable. Cluster 11 had no events, and this is represented by an empty plot. Cluster 8 looks a little sparse, with just a small purple segment representing school absence, but this highlights that cluster 8's only characteristic was school absence and that the absence levels were at a relatively low level compared to the other clusters. Overall this plot aims to provide a more intuitive visualisation of the cluster characteristics in order to complement Table 13.

Figure 28: Nightingale plot of cluster characteristics

Table 15 considers attributes that were contained in the data but were not clustered upon. The attribute detailing number of address changes was not used for clustering as it was thought that it may be unreliable data; it was derived from the address data, but this was not necessarily up to date for all families. Nevertheless, there were clear differences between clusters, with cluster 3 having almost two thirds of families changing address at least once in the year prior to intervention. The rows detailing the percentage of families with pre-existing Child Protection Plans (CPP) and Looked after Child (LAC) events were

included to highlight that although these families did not have new events pertaining to these in the year before intervention, they were already ongoing prior to this.

Table 15: Percentage of families with each event per cluster, for events not clustered on (with notable percentages highlighted in bold)

Cluster	**1** n=291	**2** n=335	**3** n=115	**4** n=61	**5** n=21	**6** n=54	**7** n=25	**8** n=223	**9** n=243	**10** n=182	**11** n=605
Receiving DWP benefits	48%	46%	35%	57%	57%	57%	**28%**	42%	36%	42%	40%
Changed address at least once	46%	53%	**73%**	49%	48%	48%	64%	**30%**	54%	42%	36%
Percentage of single person families	8%	3%	4%	10%	14%	0	12%	3%	2%	4%	**41%**
Drug/Alcohol Events	2%	5%	2%	5%	0	4%	0	1%	3%	4%	1%
Domestic Abuse Events	14%	17%	7%	5%	**33%**	6%	20%	0	13%	14%	0
Percentage with no children (aged < 18)	11%	0.3%	0	18%	**48%**	0	4%	0	1%	0	**50%**
Percentage with no adult (aged >= 18)	4%	8%	7%	5%	0	7%	8%	13%	10%	13%	8%
Pre-existing CPP	5%	10%	6%	2%	10%	2%	4%	12%	1%	3%	3%
Pre-existing LAC	1%	1%	1%	2%	0	0	4%	2%	0	0	1%

Table 16 details the treatments types for the first intervention by cluster. There were clear differences across the clusters where intervention treatment types were considered, which would seem to be logical, as intervention treatment should target different types of problems. To illustrate these differences better, Figure 29 plots them (excluding the 'Other' treatment type as this included only one family).

Table 16: First intervention treatment types by cluster (with notable percentages highlighted in bold)

Cluster	**1** n=291	**2** n=335	**3** n=115	**4** n=61	**5** n=21	**6** n=54	**7** n=25	**8** n=223	**9** n=243	**10** n=182	**11** n=605
Intervention type:											
AO	23%	10%	12%	25%	33%	13%	20%	30%	25%	20%	**35%**
CFPT	32%	25%	23%	16%	29%	13%	16%	**39%**	36%	31%	29%
FF	14%	24%	**51%**	10%	0	19%	24%	8%	9%	8%	12%
FINIS	4%	2%	2%	2%	5%	6%	0	4%	**6%**	4%	1%
FIP	27%	38%	**11%**	48%	33%	**50%**	40%	19%	24%	36%	22%
Other	0	0	0	0	0	0	0	0	0	1%	0

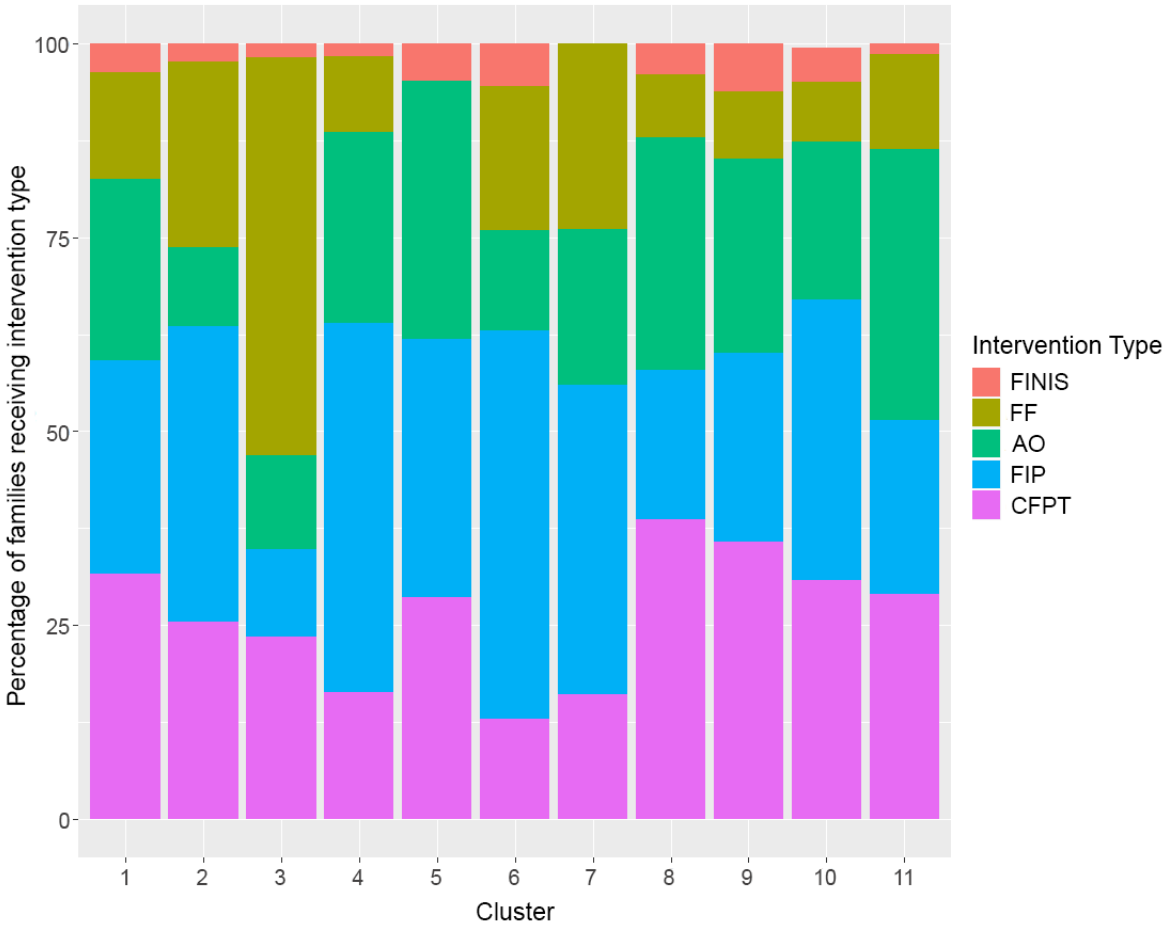


Figure 29: The percentage of families receiving each intervention type by cluster (ECC data)

The plot highlights that families in cluster 3 received proportionally more FF (Families First) treatment than families in any other cluster. Since the main characteristic of cluster 3 was that it contained families who had at least one Looked After Child (LAC) event, and the FF treatment is aimed at families with LAC problems, this would seem to make sense. Families in cluster 6 received the highest proportion of FIP (Family Intervention Project) treatment; this was aimed at families with the most complex needs, which suggests that a higher proportion of families in cluster 6 have a complex mixture of problems.

Where treatment resolution was considered (Table 17), although the percentages were broadly similar across the clusters, there was some variation. Clusters 2 and 3 had a higher percentage of planned endings than the other clusters, whereas cluster 5 had a particularly low percentage. However, the small size of cluster 5 should be considered when evaluating this value. Aside from cluster 5, clusters 6 and 8 had the lowest percentage of planned endings, this is interesting as the main characteristic of both these clusters was school absence. The main characteristic of clusters 2 and 3, which had the most planned endings was child safeguarding (CPP and LAC).

Table 17: First Intervention treatment outcomes by cluster (with notable percentages highlighted in bold)

Cluster	1 n=291	2 n=335	3 n=115	4 n=61	5 n=21	6 n=54	7 n=25	8 n=223	9 n=243	10 n=182	11 n=605
Intervention status:											
Open	7%	4%	2%	3%	0	7%	8%	8%	6%	5%	5%
Planned Ending	75%	80%	**81%**	75%	57%	70%	72%	69%	75%	72%	75%
Unplanned Ending	18%	17%	17%	21%	**43%**	22%	20%	24%	19%	23%	20.0%

Table 18 and Figure 30 detail the percentage of children in each cluster grouped by the OFSTED rating for the school they attended during the year prior to their first intervention. The final row of Table 19 also details the cumulative percentage of children who attended schools rated as 'good' or 'outstanding'. Not all children could be linked to a school OFSTED rating (some were not of school age, or else they had no school records, etc.), so the percentages include only the children that could be linked. Overall, 56% (1562 out of 2772) of children of approximate school age (aged between 5 and 16 a year before the first intervention date) could be linked to a school OFSTED rating. At least half of the school age children in all but clusters 4, 5, 9 and 11 could be linked to an OFSTED rating; 42% of children in cluster 4 were linked and for clusters 5, 9 and 11 just under a third could be linked. The top row, detailing the cluster number, also details the number of children that could be linked, to provide context.

Table 18: School OFSTED ratings by cluster, utilising ECC data linked to Department for Education (2016) data

	Cluster										
Percentage of children who attended schools rated by OFSTED as:	**1** n=316	**2** n=292	**3** n=80	**4** n=52	**5** n=6	**6** n=62	**7** n=25	**8** n=305	**9** n=80	**10** n=236	**11** n=108
Outstanding	8.9%	14.0%	23.8%	3.9%	16.7%	8.1%	12.0%	10.2%	20.0%	8.9%	14.8%
Good	52.2%	63.4%	57.5%	61.5%	50.0%	56.5%	76.0%	63.0%	57.5%	67.0%	64.8%
Requires Improvement	37.0%	18.5%	18.8%	30.8%	16.7%	33.9%	8.0%	24.3%	22.5%	19.1%	18.5%
Inadequate	1.9%	4.1%	0	3.9%	16.7%	1.6%	4.0%	2.6%	0	5.1%	1.9%
Good or Outstanding	61.1%	77.4%	81.3%	65.4%	66.7%	64.5%	88.0%	73.1%	77.5%	75.8%	79.6%

For each child, the OFSTED rating that would have applied during the year before intervention was utilised (if there was more than one); if there was no OFSTED rating available for this time period, then the rating dated most recently after the first intervention date was utilised.

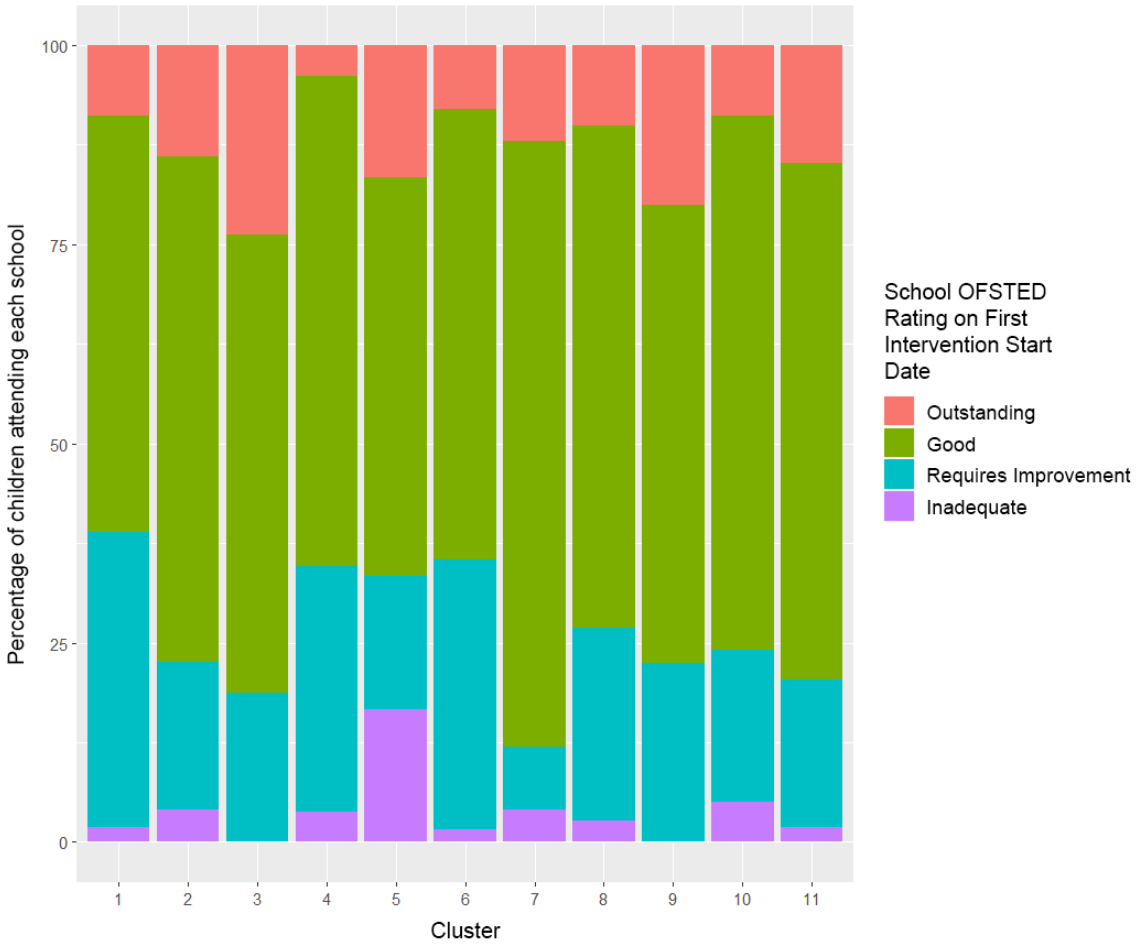


Figure 30: Percentage of children in each cluster attending schools with each OFSTED rating, utilising ECC data linked to Department for Education (2016) data

Figure 30 illustrates that, comparatively, cluster 3 had the highest percentage of children attending Outstanding schools, and cluster 4 the lowest percentage. Clusters 3 and 9 had no children attending Inadequate schools. Whilst cluster 5 did have the highest percentage of children attending Inadequate schools, there were only 6 children in Cluster 5 who linked to OFSTED data, and so this number could be misleading.

Cumulatively, cluster 1 had the lowest percentage of children attending good or better schools, whereas cluster 7 had the highest percentage. However, the small size of cluster 7 (only 25 children could be linked to an OFSTED rating) might mean that such a high percentage could simply be a quirk of the data. Clusters 3 and 11 also had higher percentages of children attending good or better schools, and they had a more reliable sample size. The school OFSTED data was analysed as it was thought that it might provide useful insight since two of the clustering attributes pertained to school events (absence and exclusion), and to some extent the OFSTED rating could also be considered a 'place-based' attribute, since children tend to attend schools that they live close to.

6.3.2 Detailed summary of clusters

There follows a detailed description of the characteristics of each of the eleven clusters, drawing together the information from the previous section.

6.3.2.1 Cluster 1: School exclusion and criminal offences

To provide a visual reminder, Figure 31, details the percentage of families in cluster 1 who had each of the particular events in the year prior to the start of first intervention. For comparison, the percentage of families who had these events for all the data (all TF) are also plotted. This comparison illustrates that cluster 1 had proportionally more families with school absence, exclusion and criminal offences, but proportionally fewer serious child safeguarding issues (CPP and LAC).

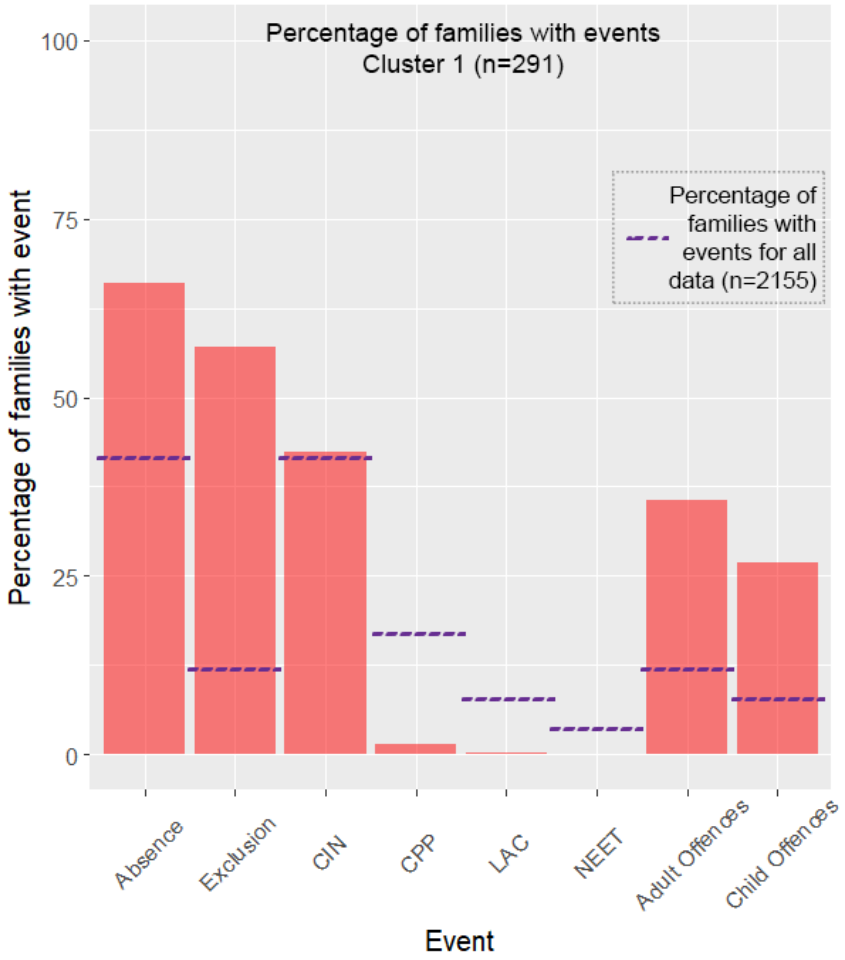


Figure 31: Percentage of families in cluster 1 with each event, with percentage of events for all families highlighted for comparison

Cluster 1 contained 291 families, consisting of 1067 individuals. Of these 602 (58%) were children, with the mean age being 10. Figure 32 plots the age distribution for adults and children and shows that a large proportion of children were in the 10 to 15 years of age range. 11% of families did not have a child (under the age of 18).

Figure 32: Age distribution of children (aged under 18 on first intervention start date) and adults in cluster 1

Two thirds (66%) of all families with school exclusion were contained in this cluster. 57% of families in cluster 1 had school exclusion (mean = 2) and this tended to be accompanied by school absence (85% of those with exclusion had absence too). However, although two thirds of families had school absence, the absence levels were low, with an average of 3.9% unauthorised school sessions per family. Only 7% of families had over 15% unauthorised absence. Of all the clusters, proportionally fewer children (61%) in cluster 1 attended a school rated 'good' or 'outstanding' by OFSTED, and cluster 1 had the highest percentage of children attending schools that required improvement.

It was also notable that just under half (47%) of all families with criminal offences committed by adults were contained in this cluster. And just under half (48%) of all families with offences committed by children were contained here too. However, families tended to have one or the other but not both; only 6% of families who had criminal offences had offences committed by both adults and children.

There were almost no serious child safeguarding events (only 4 families had Child Protection Plans, and 1 family had Looked after Children). However, 16 families had a pre-existing CPP, and 4 had previous LAC events that were likely to be ongoing (as they had no end date), but given the cluster size these were low numbers. 42% of families had CIN events; given the spread of CIN events over all clusters this was not particularly remarkable. No families had members who were logged as NEET.

This cluster had a fairly low average silhouette value (0.22) and might be considered only loosely cohesive. 11% of records/families had a negative individual silhouette value, indicating that these families might have been better suited to another cluster.

In terms of intervention treatment, 32% of families received Complex Families Parenting Treatment (CFPT), 28% were receiving Family Intervention Project (FIP), 23% were receiving Assertive Outreach (AO), and smaller percentages were receiving FF and FINIS.

6.3.2.2 Cluster 2: Child Protection

Figure 33 details the percentage of families in cluster 2 who had each of the particular events in the year prior to the start of first intervention. For comparison, the percentage of families who had these events for all the data (all families) are also plotted. The plot illustrates that cluster 2 had proportionally far more families with Child Protection Plans.

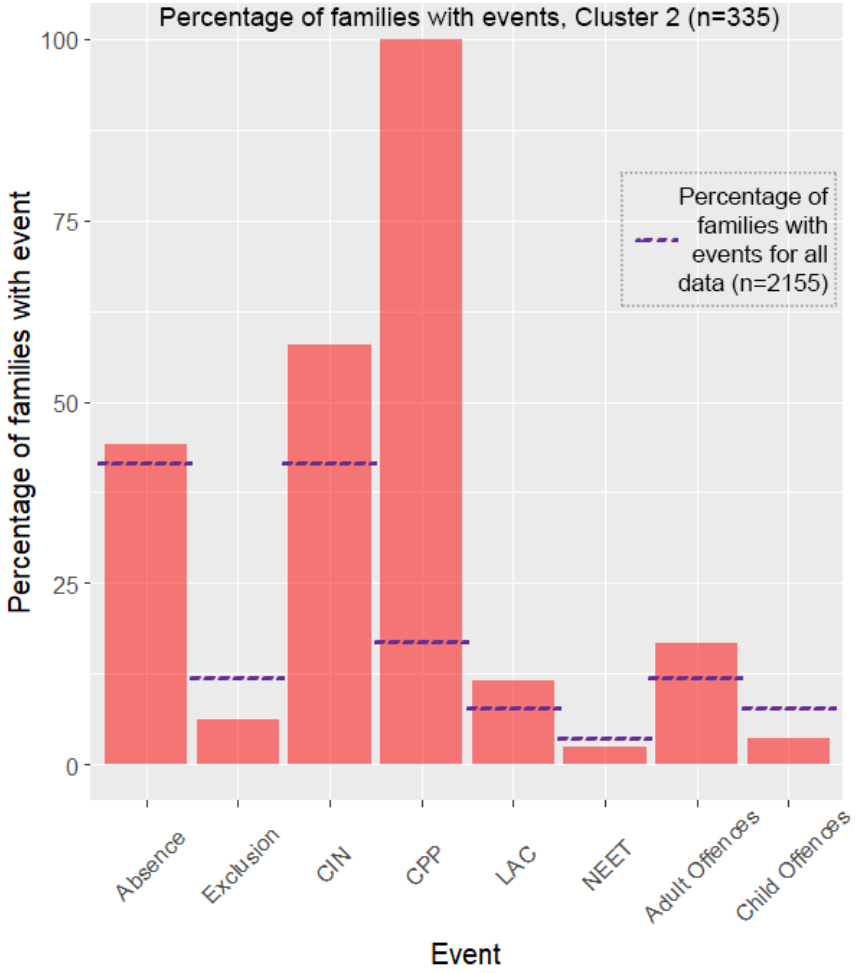


Figure 33: Percentage of families with each event for cluster 2, with percentage for all families highlighted

Cluster 2 contained 335 families, consisting of 1282 individuals. Of these 753 (59%) were children, with the mean age being 7. Figure 34 plots the age distribution for adults and children and shows that a large proportion of children were aged under 11.

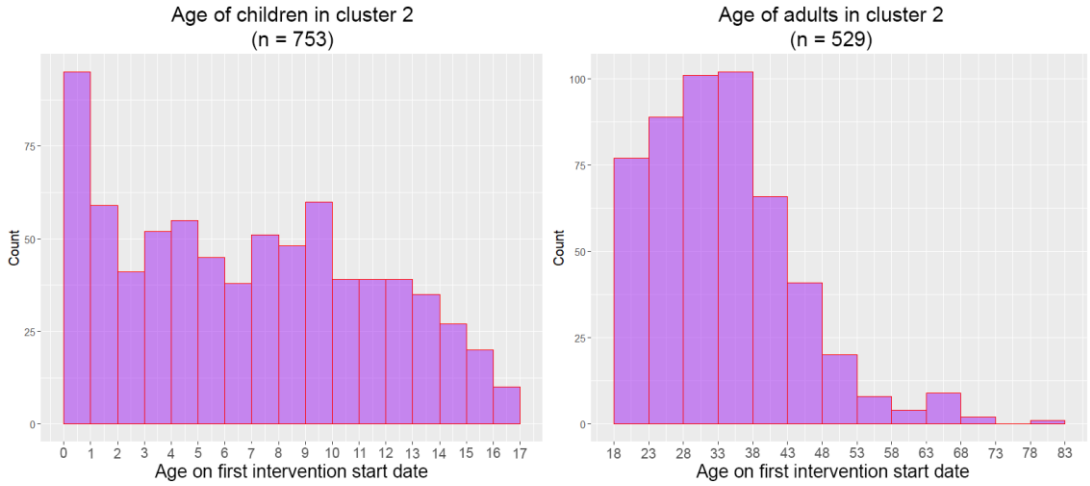


Figure 34: Age distribution of children (aged under 18 on first intervention start date) and adults in cluster 2

All families had Child Protection plans in cluster 2; 53 families (16%) had only a CPP and no other events. Of all the families with CPPs, 92% were contained in this cluster. 58% of families had CIN events, which was proportionally higher than for most of the other clusters, but this may be because CIN events are somewhat correlated with CPPs. 12% of families had LAC events, and of all the families with LAC events almost a quarter (23%) were contained in this cluster.

There was very little school exclusion; 6% of families had them, which was the lowest proportion of all the clusters (1-7). 44% of families had school absence, but the levels were low, with an average of 3.7% unauthorised sessions per family. 15% of families had over 15% unauthorised absence. 77% of children in cluster 2 attended a school rated 'good' or 'outstanding' by OFSTED; this was comparatively high.

There were very few families with members who were NEET (2%) or those with criminal offences committed by children (4%). However, 17% of families had criminal offences that were committed by adults. Of all the families with adult criminal offences, a quarter (25%) were contained in this cluster. 17% of families had events that were classed as domestic abuse.

Cluster 2 represents one of the more diverse clusters, as 44% of families had 3 or more different types of events occur in the year prior to intervention. The average silhouette width of 0.44 was acceptable, and only 4% of families had negative individual silhouette values, meaning that the majority of families most likely did belong in this cluster

In terms of Intervention treatment type, 38% of families were receiving Family Intervention Project (FIP), which aims to engage the most challenging families. 25% had

Complex Families Parenting (CFPT), and 24% Families First (FF), with smaller percentages receiving AO and FINIS. Almost all families receiving FF and FINIS had planned endings.

6.3.2.3 Cluster 3: Looked After Children

Figure 35 details the percentage of families in cluster 3 who had each of the particular events in the year prior to the start of first intervention. For comparison, the percentage of families who had these events for all the data (all families) are also plotted. The plot illustrates that comparatively families in cluster 3 had much higher proportions of LAC events, and also higher proportions of CIN events and offences committed by adults; but lower proportions of CPP events and school absence.

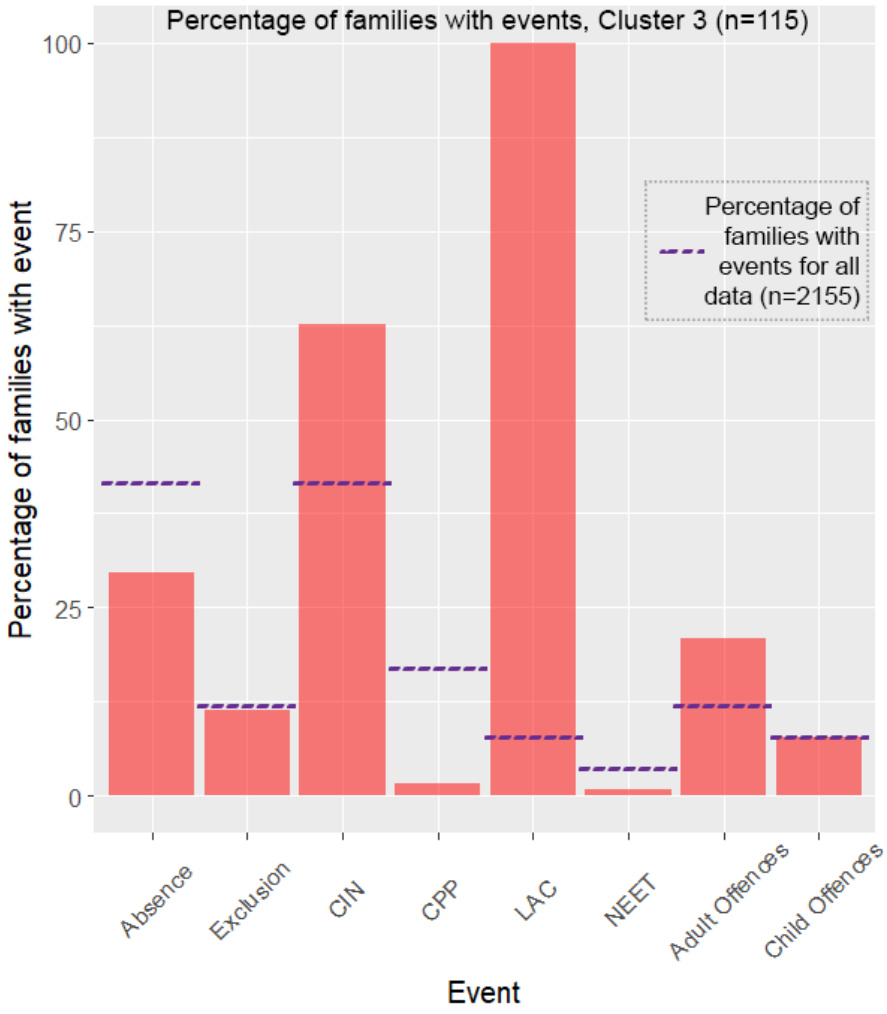


Figure 35: Percentage of families with each event for cluster 3, with percentage for all families highlighted

Cluster 3 contained 115 families, consisting of 424 individuals. Of these 233 (55%) were children, with the mean age being 8. Figure 36 plots the age distribution for adults and children and shows a mixed distribution for the children, with a peak at age 6.

Figure 36: Age distribution of children (aged under 18 on first intervention start date) and adults in cluster 3

All families in this cluster had Looked After Children events, 18% of families had only LAC events and no other event. Of all the families with LAC events, 69% were contained in this cluster. Just under two thirds of families (63%) had CIN events, which was comparatively high, but few had Child Protection Plans (2%, which rose to 8% when considering pre-existing plans).

Just under a third (30%) of families had school absence, which was proportionally lower than most of the other clusters. And the average percentage of unauthorised absence per family was 1.2%, which was a very low level; only 3% of families had school absence that was greater than 15% of available sessions. There were also low levels of school exclusions (11% of families). Cluster 3 had the highest percentage (24%) of children attending schools OFSTED rated as 'outstanding'; and 81% of children attended a school with a rating of 'good' or 'outstanding', which was the second highest percentage of all the clusters. Only one family had NEET members.

Few families had criminal offences committed by children (8%), however a comparably higher proportion had offences committed by adults (21%). It was notable that most families who had offences committed by adults did not also have school absence.

Just over a third of families (35%) were receiving DWP benefits, and this was a lower percentage than for most other clusters. 74% of families had changed address at least once in the year prior to intervention, which was a much higher percentage than any of the other clusters. However, this might be explained by the fact that all families in this cluster contain children who were moving around (in social care), therefore it is likely that address changes might be frequent.

Cluster 3 had the highest silhouette value (0.56) of all the clusters and so might be thought of as the most cohesive. Only 2 families (2%) had individual silhouette values that were negative, and which indicated that they might be better suited to another cluster.

In terms of Intervention treatment type, just over half (51%) of families in this cluster were receiving Families First (FF) treatment; this was a far higher percentage than for any other cluster. As FF works with families who are at risk of needing social care and works to try and keep them together, it would appear to make sense that so many families from this cluster should be receiving this treatment. Almost all families that had FF treatment had a planned ending.

6.3.2.4 Cluster 4: NEETs

Figure 37 details the percentage of families in cluster 4 who had each of the particular events in the year prior to the start of first intervention. For comparison, the percentage of families who had these events for all the data (all families) are also plotted. The plot highlights the comparatively high proportion of families with NEET members and criminal offences committed by children, and lower proportions of child safeguarding (CIN & CPP) events.

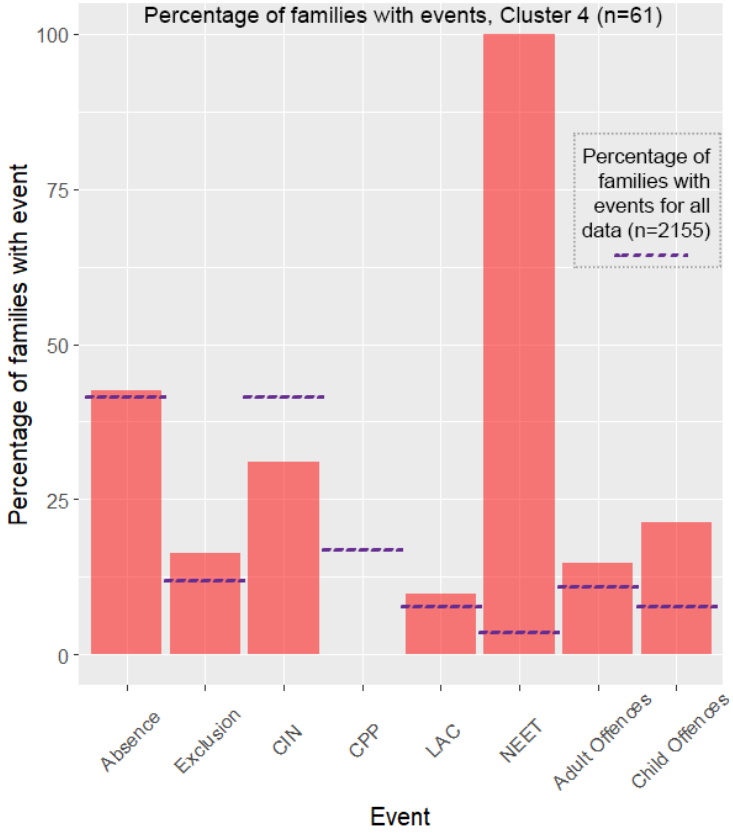


Figure 37: Percentage of families with each event for cluster 4, with percentage for all families highlighted

Cluster 4 contained 61 families, consisting of 270 individuals. Of these 157 (58%) were children, with the mean age being 11. Figure 38 plots the age distribution for adults and children, and for the children it shows a larger proportion in their teens, particularly aged 16 to 17. For the adults, there was a very large proportion in the 18-20 age range. 18% of families in this cluster contained no children (aged under 18). However, surveying the age ranges revealed that this cluster generally contained families with older children, therefore although some families contained no individuals aged under 18, the 'children' in these families were simply aged 18 or above and were classed as adults for the purpose of this analysis.

Figure 38: Age distribution of children (aged under 18 on first intervention start date) and adults in cluster 4

All families in this cluster had at least one member who was classed as NEET (not in employment, education or training). A quarter (25%) of families had only NEET members and no other events. Of all the families with NEET members just over three quarters (78%) were contained in this cluster.

This cluster contained low levels of child safeguarding: only one family had a pre-existing CPP; 11% had LAC events; and 31% had CINs (which was a lower percentage than all but one cluster). There were also fairly low levels of school absence and exclusion. 44% of families had some school absence, and the average percentage of unauthorised absence per family was low, at 2.2%. Only 3% of families had school absence that was greater than 15% of available sessions. 16% of families had school exclusions. The percentage of children attending schools judged as 'good' or 'outstanding' by OFSTED was 65%, which was lower than all but two of the clusters; overall, cluster 4 had the lowest percentage (4%) of children attending 'outstanding' schools.

Cluster 4 had an acceptable silhouette value (0.44) and appears fairly cohesive. However, 5 families (8%) had negative individual silhouette values indicating that they might be better suited to another cluster.

In terms of Intervention treatment type, almost half (48%) of families in this cluster were receiving Family Intervention Project (FIP) treatment, which is aimed at the most challenging families.

6.3.2.5 Cluster 5: Adult Criminal Offences

Figure 39 details the percentage of families in cluster 5 who had each of the particular events in the year prior to the start of first intervention. For comparison, the percentage of families who had these events for all the data (all families) are also plotted. The plot illustrates that cluster 5 had comparatively high proportions of families with criminal offences committed by adults, and school exclusion; conversely, child safeguarding and school absence events were proportionately low.

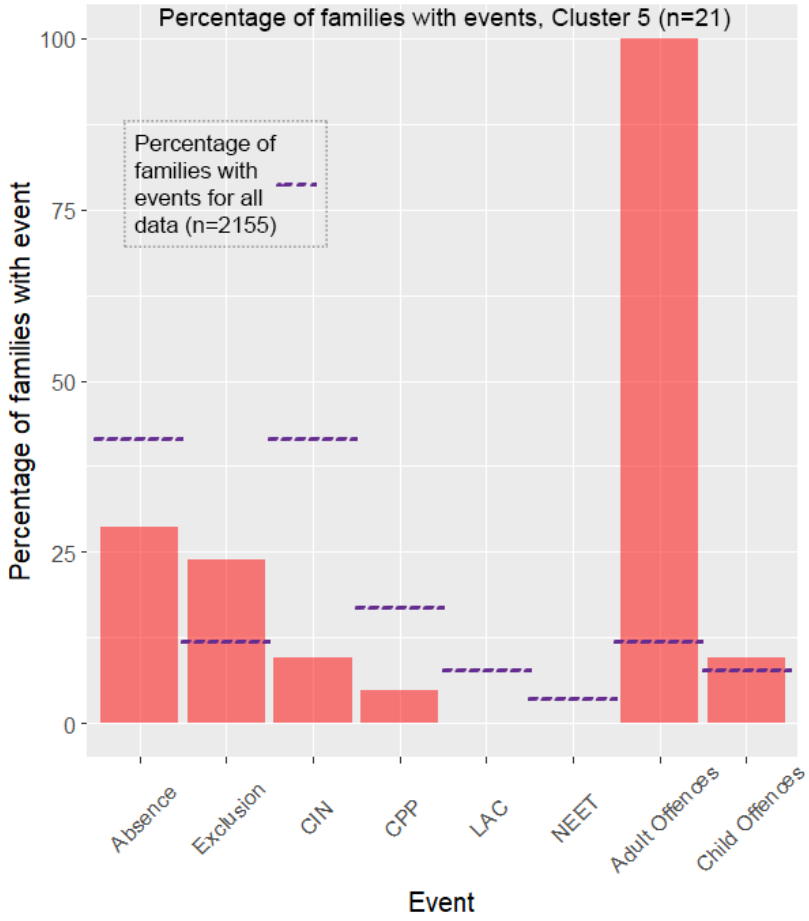


Figure 39: Percentage of families with each event for cluster 5, with percentage for all families highlighted

Cluster 5 contained 21 families, consisting of 70 individuals. Of these 27 (39%) were children, with the mean age being 10. Figure 40 plots the age distribution for adults and

children, and for the children it shows a large proportion in their teens, aged 13 and up. This was a particularly small cluster and so it was not surprising that there were gaps in the age distributions. Just under half (48%) of all families in this cluster had no children (aged under 18).

Figure 40: Age distribution of children (aged under 18 on first intervention start date) and adults in cluster 5

All families in this cluster had at least one adult who had committed a criminal offence, and the levels of criminal offences were high, with a mean of 4 per family. A third of families (33%) had domestic abuse events, which was the highest proportion over all clusters. The families contained in this cluster did not have a diverse mix of events; almost two thirds (62%) of families had only criminal offences committed by adults and no other events.

Proportionally fewer families had school absence (29%) than most of the other clusters; this might be explained by the fact that just under half of the families had no children. However, those families that did have school absence had high levels of it; half had more than 15% unauthorised sessions. And those families with school absence tended to also have school exclusion; just under a quarter (24%) of families had school exclusion. The percentage of children attending schools judged as 'inadequate' by OFSTED was 17%, the highest for any cluster; conversely, 67% attended schools judged as 'good' or 'outstanding' which was fairly low in comparison to the other clusters. However, only 6 children (of a possible 27) could be linked to OFSTED data, therefore, this sample is so low that the statistics may be unreliable.

There were generally low levels of child safeguarding; this again might be explained by the fact that there were fewer children in this cluster. 10% of families had children who had committed criminal offences. There were no NEET members.

The silhouette value of this cluster (0.46) was the second highest of all clusters, and none of the families had a negative individual silhouette value, indicating that all families were probably best suited to this cluster. In terms of Intervention treatment type, most families received either AO (33%), CFPT (29%) or FIP (33%)

6.3.2.6 Cluster 6: High Levels of School Absence

Figure 41 details the percentage of families in cluster 6 who had each of the particular events in the year prior to the start of first intervention. For comparison, the percentage of families who had these events for all the data (all families) are also plotted. The plot indicates that, comparatively the families in cluster 6 had higher proportions of all events, most notably school absence, school exclusion and criminal offences committed by children.

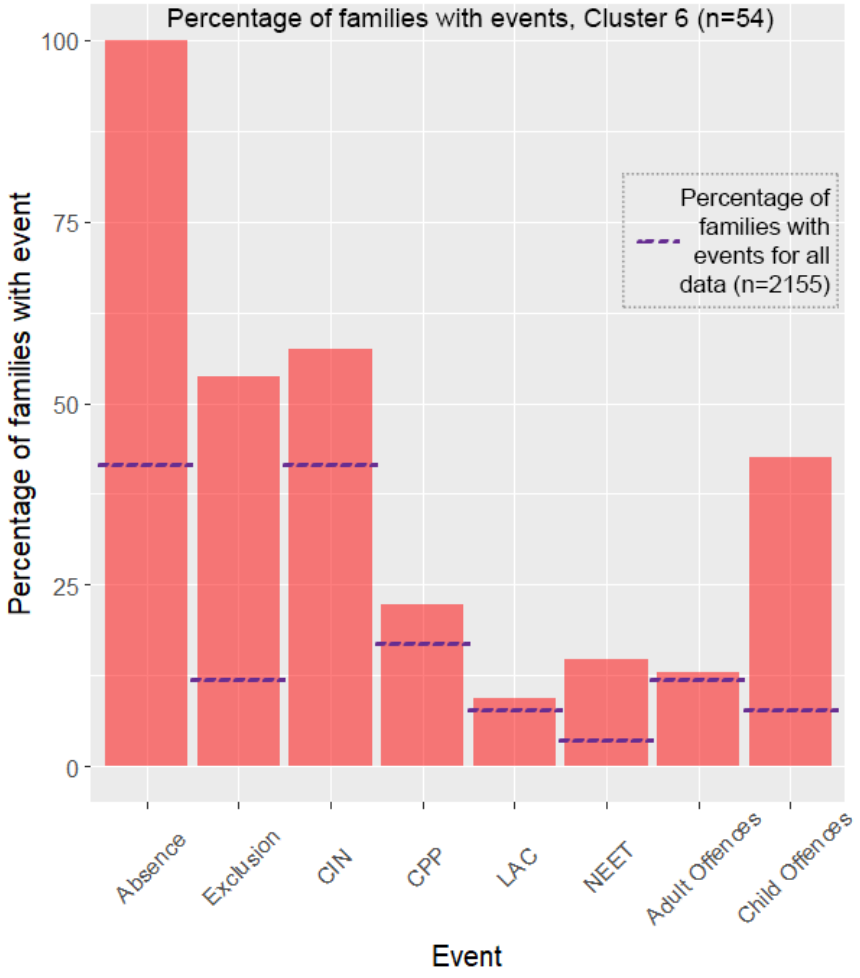


Figure 41: Percentage of families with each event for cluster 6, with percentage for all families highlighted

Cluster 6 contained 54 families, consisting of 206 individuals. Of these 123 (60%) were children, with the mean age being 12. Figure 42 plots the age distribution for adults and children, and for the children it shows a large proportion in their teens, particularly aged 14 to 15. For the adults, there was a large proportion in the 18-20 range and also mid to late thirties. All families had children, although 7% had no adults attached to them.

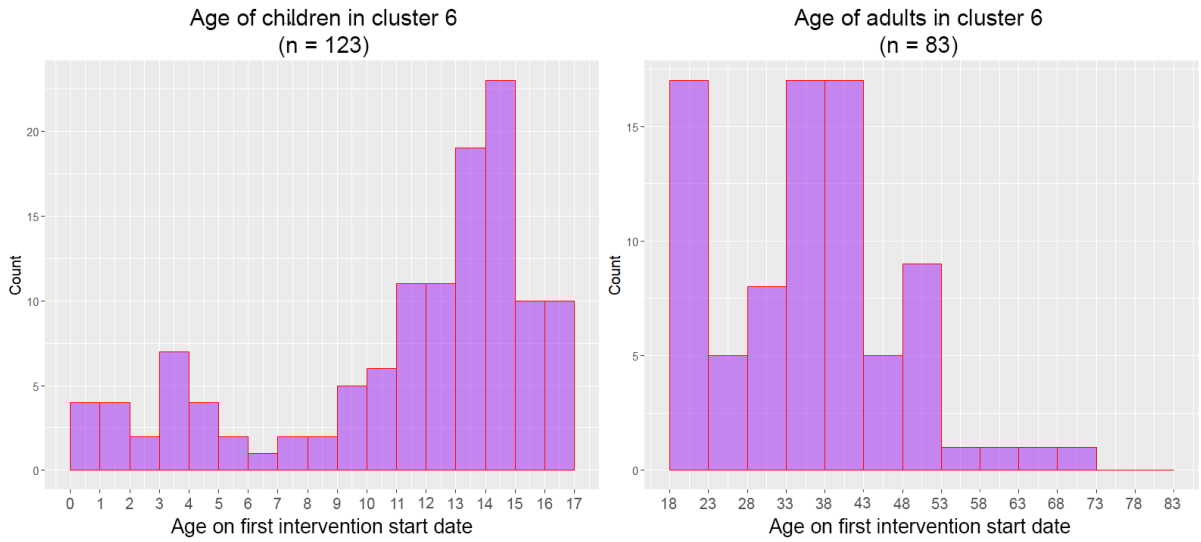


Figure 42: Age distribution of children (aged under 18 on first intervention start date) and adults in cluster 6

This cluster contains families who all had high levels of school absence (mean 39%). All but one family (98%) had unauthorised school absence greater than 15%. This was a diverse cluster and all families had school absence combined with at least one other issue; 78% had three events or more. Just over half of families (54%) had school exclusion, which was the second highest level of all clusters. The percentage of children attending schools judged as 'good' or 'outstanding' by OFSTED was 65%, which was the second lowest percentage in comparison to the other clusters.

There were proportionally higher levels of child safeguarding events than for most of the other clusters, with 57% having CIN events, 24% having CPPs and 9% having LAC events. In terms of crime, 48% of families had criminal offences committed by children, which was comparatively high, and 13% had criminal offences committed by adults. 15% of families had at least one NEET member, which was also a comparatively high proportion.

The silhouette value of this cluster (0.14) was the lowest of all clusters, and 19% of the families had a negative individual silhouette value, indicating that they may have been better suited to another cluster. This implies that this cluster was not very cohesive, and any conclusions drawn from it should be treated cautiously.

In terms of Intervention treatment type, half of families (50%) of families in this cluster were receiving Family Intervention Project (FIP) treatment; this was the highest proportion over all clusters. This treatment is aimed at the most challenging families who have complex needs.

6.3.2.7 Cluster 7: Child Criminal Offences

Figure 43 details the percentage of families in cluster 7 who had each of the particular events in the year prior to the start of first intervention. For comparison, the percentage of families who had these events for all the data (all families) are also plotted. The plot illustrates that families in cluster 7 had comparatively high proportions of children who committed criminal offences, and child protection plans.

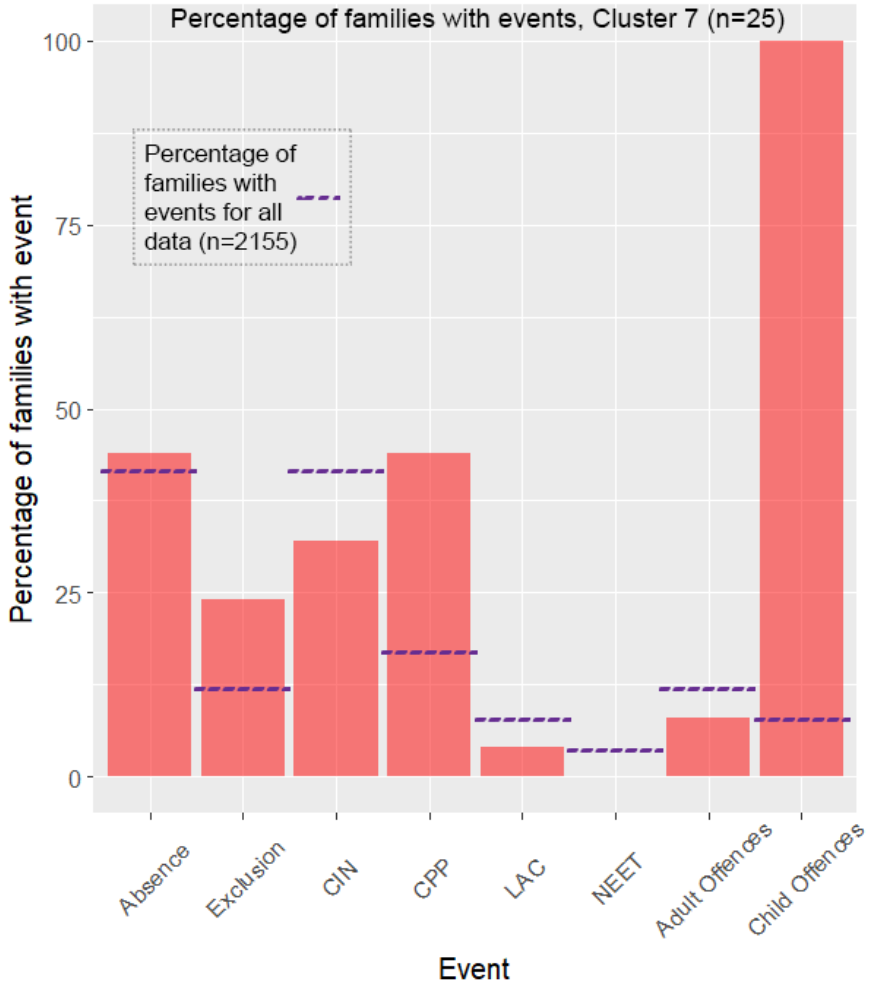


Figure 43: Percentage of families with each event for cluster 7, with percentage for all families highlighted

Cluster 7 contained 25 families, consisting of 86 individuals. Of these 53 (62%) were children, with the mean age being 12. Figure 44 plots the age distribution for adults and children, and for the children it shows a larger proportion in their teens, aged 13 and up. 8% of families had no adult attached to them.

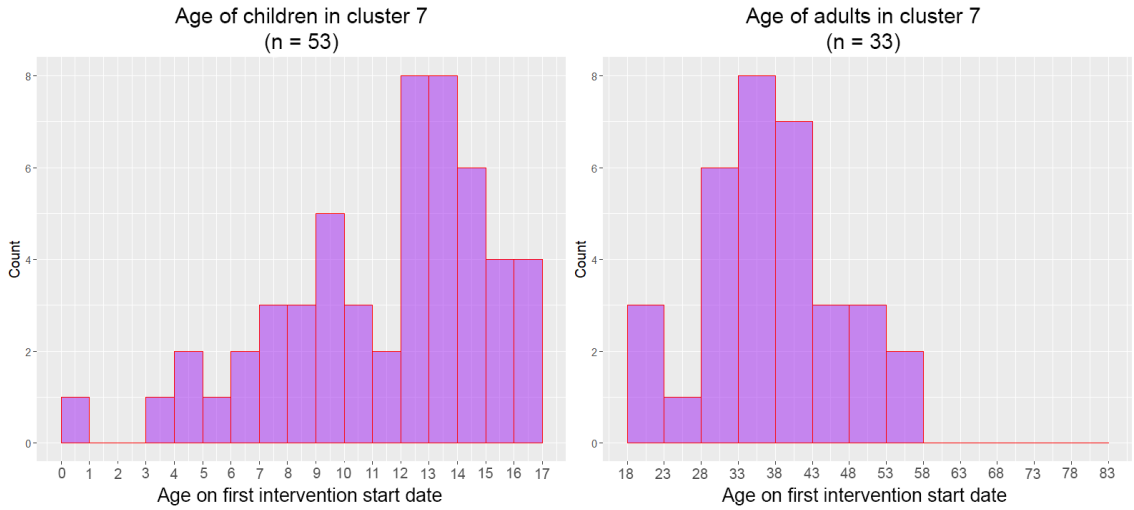


Figure 44: Age distribution of children (aged under 18 on first intervention start date) and adults in cluster 7

This cluster contained families who all had at least one child who had committed criminal offences, with the mean number of offences being 4 per family. Just over a quarter (28%) of families had only criminal offences committed by children and no other event. Only 8% of families had criminal offences that were committed by adults.

Just under half of all families (44%) had school absence, although the absence levels were fairly low, with an average of 2.9% unauthorised absence. Just under a quarter of families (24%) had school exclusion. There were no NEET members. The percentage of children attending schools judged as 'good' or 'outstanding' by OFSTED was 88%, which was the highest percentage in comparison to the other clusters. However, given the small size of this cluster this could be a quirk of the data.

There were high levels of child safeguarding. Just under half (44%) of families had CPPs, which is the second highest proportion of all the clusters. Those families with CPPs also tended to have school absence. Just under a third of families had CIN events; families with CIN events tended not to have CPPs. 8% of families had LAC events. A fifth (20%) of families had events that were considered domestic abuse. Just over a quarter (28%) of families were receiving DWP benefits, which was proportionally lower than for all the other clusters.

This is a particularly small cluster with only 25 families. The average silhouette value was fairly low (0.2) and 16% of families had a negative individual silhouette value, indicating that they might be better suited to another cluster. This indicates a possible lack of cohesion and any conclusions drawn from this cluster must be treated with caution.

In terms of Intervention treatment type, families were receiving either AO (20%), CFPT (16%), FF (24%) or FIP (40%).

6.3.2.8 Cluster 8: School Absence Only

Cluster 8 contained 223 families, consisting of 806 individuals. Of these 525 (65%) were children, with the mean age being 9. Figure 45 plots the age distribution for adults and children, and for the children it shows a larger proportion aged 8 to 12 years. All families had children, as would be expected, but 13% of families in this cluster had no adult attached to them. Whilst all but cluster 5 had a small percentage of families with no adults, 13% was the highest percentage over all clusters (cluster 10 also had 13% of families with no adults).

Figure 45: Age distribution of children (aged under 18 on first intervention start date) and adults in cluster 8

All families in this cluster had school absence and no other events in the year prior to intervention. The average percentage of unauthorised school sessions was 6.4%, and only 12% of families had unauthorised absence that was greater than 15%. The percentage of children from this cluster attending schools judged as 'good' or 'outstanding' by OFSTED was 73%, which was comparatively lower than for all but four of the clusters.

Although these families had no events other than school absence in the year prior to intervention, some families had pre-existing child safeguarding issues. 12% of families had ongoing CPPs, and 2% of families had previous LAC events, however they had no logged events in the year prior to intervention.

Just under a third (30%) of families had at least one address change in the year prior to intervention. Comparatively this was a lower percentage than for any of the other clusters. It may be that these families do not move as frequently as others, however it may also be possible that if a family does not have any of the other events (CIN, criminal offences, etc.), then address changes are less likely to be logged.

In terms of Intervention treatment type, the majority of families were receiving AO (30%), FIP (19%) or CFPT (39%). In comparison to the other clusters, this cluster had the highest proportion of families receiving CFPT. This treatment provides parenting interventions (lessons and support) to families with a range of complex needs. Since, these families only had school absence, the treatment type could indicate the possibility that at least some of these families might have had other needs there were simply not represented by the available data.

6.3.2.9 Cluster 9: Children in Need only

Cluster 9 contained 243 families, consisting of 834 individuals. Of these 486 (58%) were children, with the mean age being 7. Figure 46 plots the age distribution for adults and children, and for the children it shows a large proportion aged 1 to 6. 10% of families in this cluster had no adult attached to them.

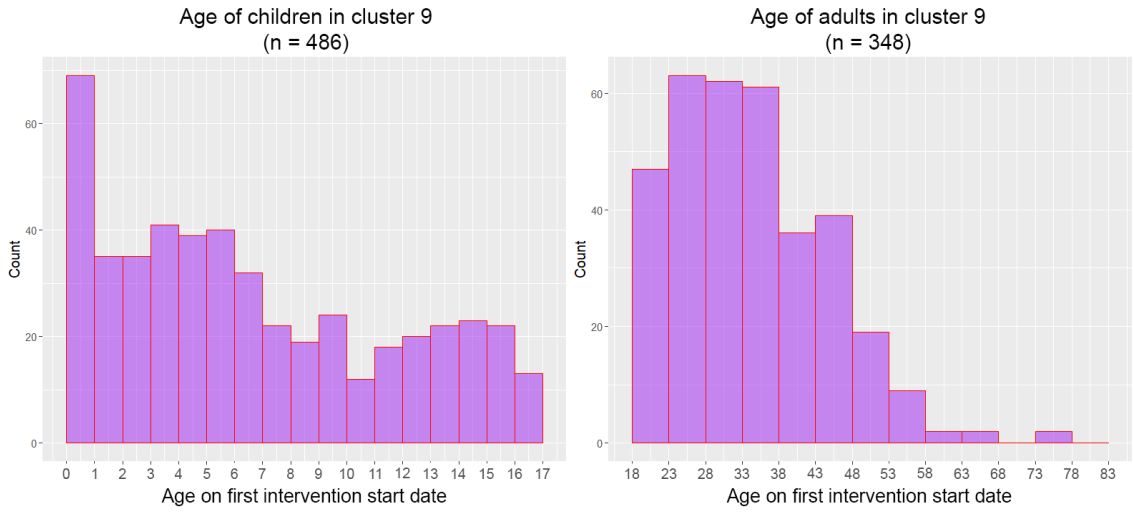


Figure 46: Age distribution of children (aged under 18 on first intervention start date) and adults in cluster 9

All families in this cluster had at least one CIN event in the year prior to intervention, but no other events. Three families (1%) had pre-existing Child Protection Plans, but there was no change to their status in the year prior to intervention. The percentage of children from this cluster attending schools judged as 'outstanding' by OFSTED was 20%,

which was comparatively high. However, a third of all children in this cluster were too young to attend school.

Just over a third (36%) of families were receiving DWP benefits on the first intervention date. This was low compared to the other clusters. In terms of Intervention treatment type, families were generally receiving AO, CFPT or FIP. However, in comparison to the other clusters, this cluster contained the largest percentage (6.2%) of families receiving Family In Need Intervention Service (FINIS). This treatment specifically targets families with Children in Need events, so this would appear to make sense.

6.3.2.10 Cluster 10: School Absence and CIN

Cluster 10 contained 182 families, consisting of 716 individuals. Of these 470 (66%) were children, with the mean age being 9. Figure 47 plots the age distribution for adults and children, and for the children there were higher proportions of children of school age (5-15). All families had children, however 13% of families in this cluster had no adult.

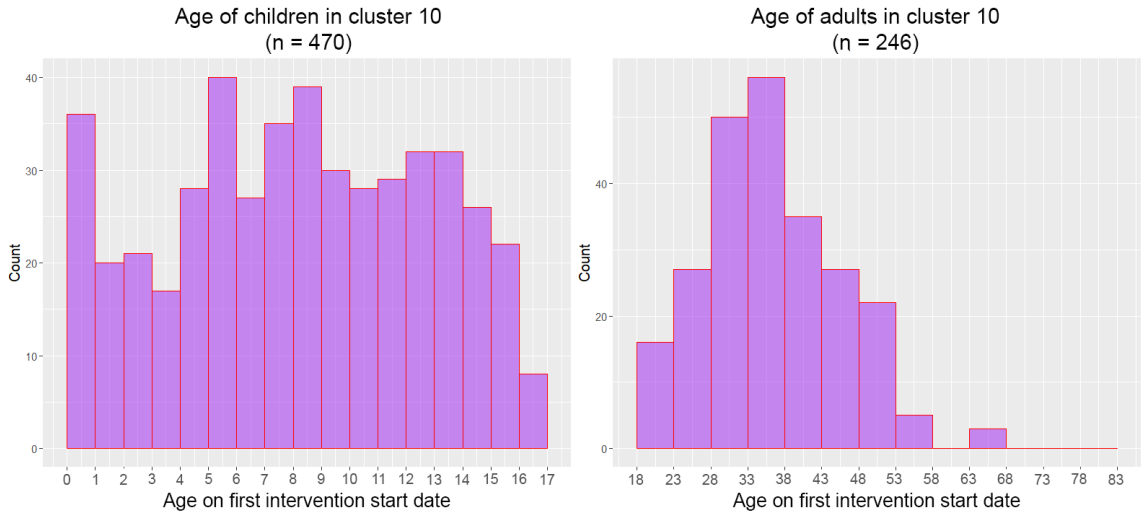


Figure 47: Age distribution of children (aged under 18 on first intervention start date) and adults in cluster 10

All families in this cluster had unauthorised school absence and at least one CIN event, but no other events in the year prior to intervention. However, 6 families (3%) had pre-existing Child Protection Plans that were issued more than a year before intervention. Aside from cluster 6, the highest levels of school absence were contained in this cluster; families had on average 10.6% unauthorised school sessions. A fifth of families had greater than 15% unauthorised school sessions. The percentage of children from this cluster attending schools judged as 'good' or 'outstanding' by OFSTED was 76%, and comparatively four clusters had a higher percentage than this.

In terms of Intervention treatment type, just over a third (36%) of families in this cluster were receiving Family Intervention Project (FIP) treatment, and 31% were receiving CFPT. These both work with families who have challenging and complex needs.

6.3.2.11 Cluster 11: No Events

Cluster 11 contained 605 families, consisting of 1333 individuals. Of these 541 (41%) were children, with the mean age being 8. Figure 48 plots the age distribution for adults and children, and for the children it is fairly even, although the plot tails away for teenagers. This was the largest cluster, and half of families (50%) had no children attached to them. In other clusters this lack of children was at least partially explained by the fact that the 'children' were older and aged 18-20 (and therefore classed as adults), however, this does not appear to be the case overall here. 8% of families had no adult attached to them.

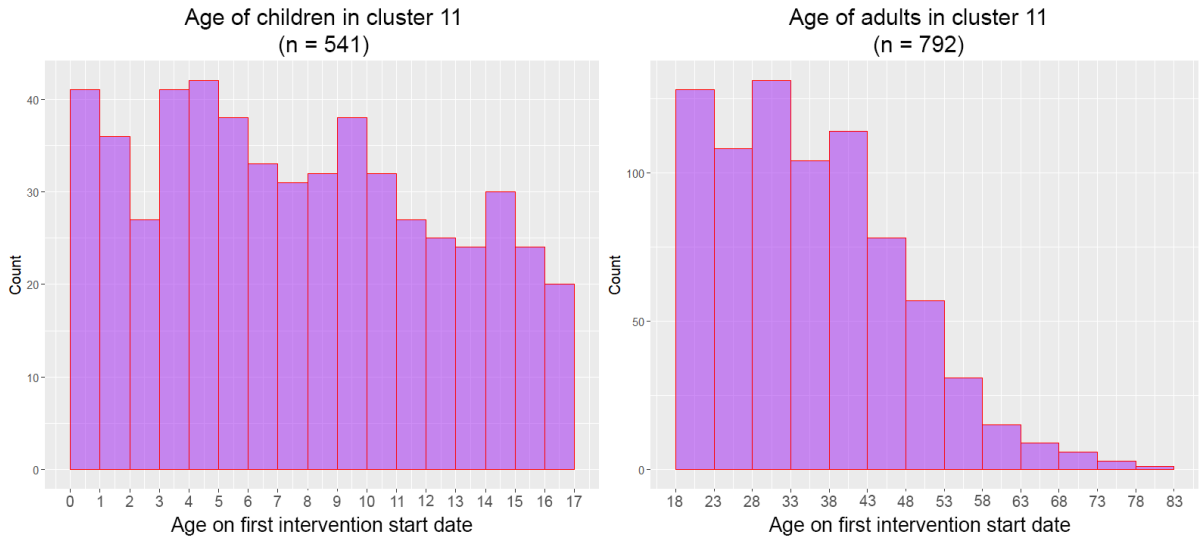


Figure 48: Age distribution of children (aged under 18 on first intervention start date) and adults in cluster 11

All families in this cluster had no events in the year prior to intervention. However, 20 families (3%) had pre-existing Child Protection Plans, 6 families (1%) had pre-existing LAC events and 9 families had members who were classed as NEET more than a year earlier. But there were no events logged pertaining to these in the year prior to intervention.

As mentioned in previous sections, the most notable aspect of this cluster was the fact that 41% of families consisted of only one person. This was a far higher percentage of single person families than for any of the other clusters (cluster 5 had 14%, and all others had less than this). The majority of these single person families consisted of lone adults, however, 10% were lone children. One possible reason for the high proportion of single

person families may be problems with the data; it is possible that some of these individuals may be part of other families, and because of errors in the data were not linked together.

The percentage of children from this cluster attending schools judged as 'good' or 'outstanding' by OFSTED was 80%, which was high in comparison to the other clusters (only clusters 3 and 7 had higher percentages). Just over a third of families (36%) had at least one address change in the year prior to intervention; this was a fairly low percentage in comparison to other clusters. However, given that these families had no events occur in the year prior to intervention, it is possible there may have been less opportunity to log address changes. 40% of families were receiving DWP benefits; this was a little lower than for the other clusters (only clusters 3 and 7 had a lower percentage).

In terms of Intervention treatment type, just over a third (35%) of families in this cluster were receiving Assertive Outreach (AO) treatment; this was the highest proportion over all clusters. As AO works with families who are at risk of developing complex needs, their treatment type may shed some light on the fact that these families appear to have no issues. It is possible that at least some of the families may have been identified as likely to develop problems in future; this may have been performed using data that was not available in this analysis (such as health or anti-social behaviour data).

6.3.2.12 Summary

Eleven different clusters of families were discovered in the data. Whilst the four pre-specified clusters (8 to 11) were fairly simple to understand, at least in terms of the events that had occurred prior to intervention (for example, just families with school absence, or families who had no events at all), the other seven clusters were more complex. However, they all had particular characteristics that were unique to the cluster and do seem to form distinct groups.

Whilst it can be difficult to assimilate such a large of body of data for each of the eleven clusters, analysing the data separately on the cluster-level, as opposed to one large analysis of the occurrence of events for all families (as in Table 6) provided far more context and a much greater understanding of the types of families that exist within the data.

However, in order to provide further clarity as to the cluster assignments, and to determine what the important factors were, it was felt that decision tree learning could provide some insight. The following analysis uses decision tree learning to derive rules for the clusters assignments and aid in the understanding of which attributes might be more important in terms of cluster assignment.

6.3.3 Using Decision Tree Learning to describe the clusters

As an alternative method of describing the cluster assignments, decision tree learning was utilised. This method was chosen to offer a visual interpretation of the cluster rules. And it was also hoped that the rules might provide further clarity around the underlying relationships within the various clusters and identify important attributes.

The CART algorithm was implemented using the R programming language and the 'rpart' package (Therneau et al., 2017). Only the data that was clustered was used in the model; that is, the data forming clusters 1 to 7, for 902 families. It was felt that including the four pre-specified clusters (8 to 11) might inflate the accuracy of the model, since they represent very simple rules. However, for completeness they were included in the overall tree plot, Figure 49.

The cluster assignment (numbered 1 to 7) was the target attribute; the eight clustering attributes (absence, offences, etc.) were the predictor attributes. The data was randomly split into a training (n = 632) and testing (n = 270) dataset, with a 70:30 split, and this was done proportionately to the target attribute. The classification model was built upon the training dataset using 10-fold cross-validation. Since the cross-validated error rate was used to decide where to prune the tree, the final model was then evaluated on the test dataset (which was not used at all in the model building process, and therefore unbiased).

The tree was pruned using a CP value of 0.0026, at the point where the cross-validated error rate was lowest. This produced a tree with 93.3% classification accuracy on the test data set. Table 19 contains a confusion matrix detailing the predicted values compared to the actual cluster assignments; the diagonal row contains the number of cases where the predicted and actual cluster assignment matched. Of the 270 records in the test dataset, only 18 were assigned to the wrong cluster (and not in the diagonal row), resulting in a 6.7% error rate. Clusters 3 and 7 were predicted with 100% accuracy; with sensitivity equal to 1, that is, all records in those clusters were correctly identified. Clusters 1 and 5

had specificity of 1, meaning that no record was wrongly identified as belonging to these clusters.

Table 19: Confusion matrix for predicted cluster assignments

		Actual Cluster Assignment						
		1	**2**	**3**	**4**	**5**	**6**	**7**
Predicted Cluster Assignment	**1**	83	0	0	0	0	0	0
	2	1	98	0	0	0	4	0
	3	0	0	35	4	0	1	0
	4	0	0	0	14	0	1	0
	5	0	0	0	0	5	0	0
	6	2	0	0	0	1	10	0
	7	1	3	0	0	0	0	7

The balanced accuracy (that is, the sensitivity plus the specificity divided by 2) for all cluster predictions was very high (greater than 0.97) for all clusters except for cluster 4 (0.89), cluster 5 (0.92) and cluster 6 (0.81). A value of 1 would mean perfect accuracy, whereas a value of 0 would mean that all predictions were incorrect, so values close to 1 indicate very high accuracy. The slightly lower accuracy for cluster 6 might be explained by the fact that cluster 6 was the least cohesive cluster, and contained some families that were probably not well fitted to it (i.e. that had negative silhouette values); therefore, in this case the rules might have been more difficult to define.

Figure 49 plots the decision tree. The extra rules for the other four pre-specified clusters were also included on the right of the tree for completeness. The plot highlights that, despite the seemingly complex nature of the clusters (as evidenced by the detailed descriptions in the previous section), the decision tree produced is fairly simple and easy to understand. It provides clear rules and in doing so highlights the key characteristics of each cluster.

Aside from cluster 7, each cluster is defined by a single rule. For instance, records are assigned to cluster 2 if the family has a child protection plan (CPP) and has fewer than two (i.e. one or none) criminal offences committed by children. Cluster 7 was defined by two rules (had two leaf nodes), suggesting this cluster was a little more complex to describe.

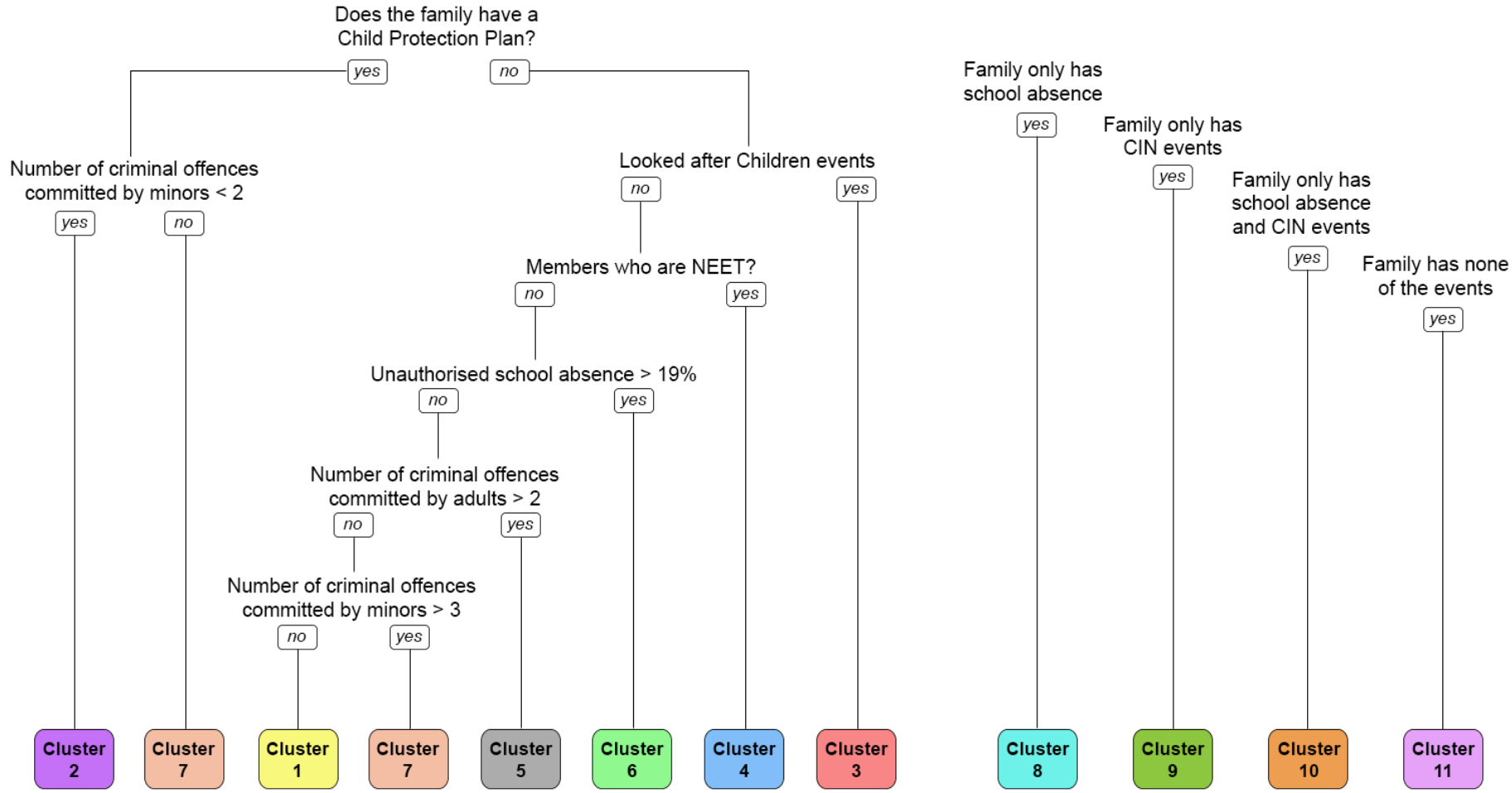


Figure 49: Decision tree visualising cluster rules, derived using the 'rpart' R package implementation of the CART algorithm and plotted with the 'rpart.plot' R package

As part of the CART process, each attribute was ranked in terms of variable importance (Table 20). This is calculated for all attributes by considering the sum of the goodness of split for each attribute in its role as either a primary or a surrogate splitter. The rpart algorithm scales these to sum to 100, with the highest score indicating the most important attribute/s.

Table 20: Variable importance scores for the decision tree predicting cluster assignment

Attribute	CPP	LAC	NEET	School Absence	Adult Offences	Child Offences	School Exclusion	CIN
Variable Importance Score	43	21	14	7	7	5	3	0

The CPP (child protection plan) attribute had the highest variable importance score, and this was reflected by it forming the main split in the tree (Figure 49). Interestingly, the CIN attribute had no importance; including it in the model (or not) made no difference to its performance. This may be because CIN events were spread across all seven clusters and therefore not unusual to any cluster. Another reason may be that many families with CIN events were not included in the decision tree (they were contained in clusters 9 and 10). School exclusion also had a low variable importance score; however, whilst not utilised as a splitter, this did have some importance as a surrogate.

One likely reason that the CPP attribute was the most important is that cluster 2 was the largest cluster (it contained 335 families) and its main feature was that all families had Child Protection Plans, therefore it would make sense that the first split in the tree would attempt to split the largest group off. However, not all families with CPPs were contained in cluster 2, therefore the next split trims off the families that belonged to cluster 7.

Figure 49 highlights that decision tree learning can be utilised to identify rules and make sense of clustering results with a high level of accuracy. It also provides an alternative (and perhaps more logical way) to think about the cluster assignments and how each individual family was assigned to its cluster. The variable importance scores are also useful in providing insight into the more important features when considering the clusters. It should be noted that in this case the decision tree produced was fairly simple and this may be because the clusters were quite well defined with each having their own particular unique characteristics. In the case of a more complex cluster analysis, a method such as this might not have such interpretable results (or may at least produce a much more complex tree).

The decision tree also provides a method of assigning new families to these clusters, if this were to be deemed appropriate in the future. For instance, if it was felt that assigning future families to the cluster most suited to them could aid in understanding a family's particular needs, then their data could be fed into the decision tree and the cluster assignment easily determined.

6.3.4 Geographical Links to Families and Clusters

All but nine of the TF (2146 out of 2155) could be linked to a Post Code; the nine families without a geographical location were excluded from the geographical analysis. For the remaining data, the post code of each family was linked to the Output Area (OA) and Lower Super Output Area (LSOA). Demographic data from the 2011 Census was then linked via the OA to each family. It was therefore possible to analyse the characteristics of the areas that each family lived in. This was important because the ECC felt that where a family lived might be a factor in whether a family was classed as a TF and perhaps also in any outcome of their treatment.

Although the ECC data contained information about the events that had happened to a family, there was no data pertaining to demographic details such as ethnicity, qualifications, religion, etc. Whilst linking each family (to the Census 2011 data for the area they lived in) cannot indicate anything about the family specifically, it can indicate that a family lived in, for example, an area with higher levels of unemployment, or an area with lower levels of people with no qualifications. It provides 'place-based' context for each family. Figure 50 plots the concentration of TF living in each LSOA, by cluster. That is, for each cluster, it plots the TF living in each LSOA as a percentage of the total number of households living there. For example, if a particular LSOA had two TF from cluster one living in it, and there was a total of 400 households altogether in that LSOA, then the percentage would be 0.5% (2 divided by 400). The number of households in each LSOA was obtained from the 2011 Census data. As already highlighted in Figure 8, there were particular areas of the city that contained higher percentages of TF generally, however Figure 50 highlights that there were subtle differences in TF location where the different cluster assignments were taken into consideration. The smaller clusters (4 to 7) do have smaller percentages, as might be expected and so are represented by paler colours, nevertheless for all clusters there were particular LSOAs that had higher proportions of TF, and these tended to vary by cluster.

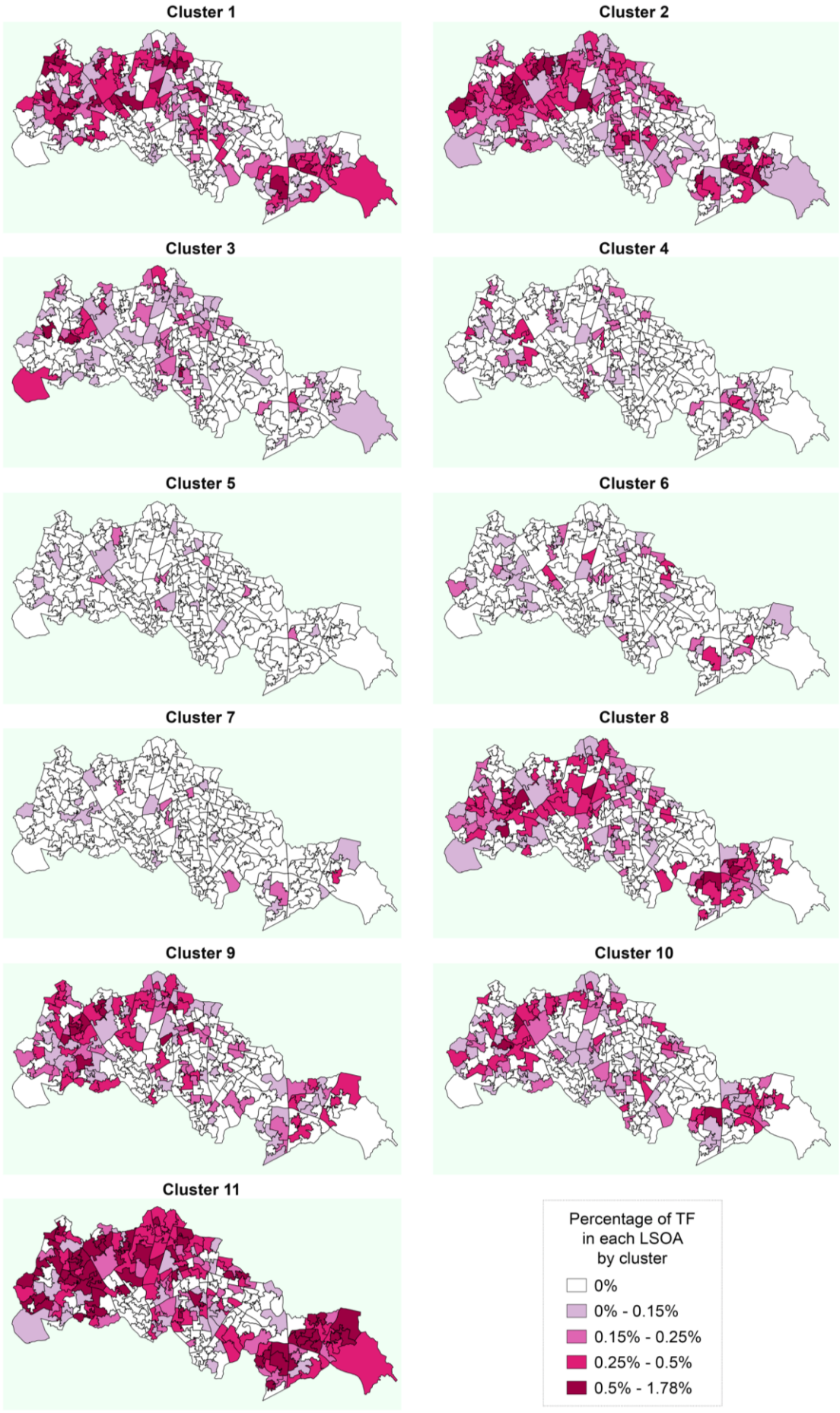


Figure 50: TF living in each LSOA as a percentage of all households in each LSOA, by cluster assignment (utilising ECC data linked to Census 2011 data)

As an alternative view Figure 51 plots heatmaps of the concentration of TF by cluster. For the purpose of comparison, no legend was included, as the concentration was different for each plot (in order to account for the different cluster sizes); the purpose of the plots was simply to provide a visual comparison of the areas with the greatest density of families for each cluster. In each plot, the highest density of families was contained in the blue areas, with smaller densities in the yellow areas and red marking the edges. The plots highlight the small size of clusters 4 to 7, with cluster 5 in particular represented simply by a circle for each family. However, it also reinforces Figure 50 which highlighted the subtle differences in the location of families from each cluster.

As a final indication of the location of TF, Figure 52 considers only TF and not the general population. For each cluster, it plots the TF living in each LSOA as a percentage of the total number of TF in that LSOA. For instance, if a particular LSOA contained ten TF, and five were from cluster 11, then the cluster 11 plot would have a percentage of 50% for that particular LSOA. As indicated in the other plots, for each cluster there appear to be unique areas that have higher proportions of TF from particular clusters.

Each of the plots does not contain any identifying geography (that is, there is no underlying map included). However, this meant that direct comparisons between maps could be difficult, therefore the heatmap plots in Figure 51 each contain an identical grid in order to allow a direct comparison of the locations. Since Figures 50 and 52 contain the LSOA boundaries, no grid was required, and the boundaries allowed for easy comparison.

Figure 51: Heatmaps of TF geographical concentration for each cluster (utilising ECC data)

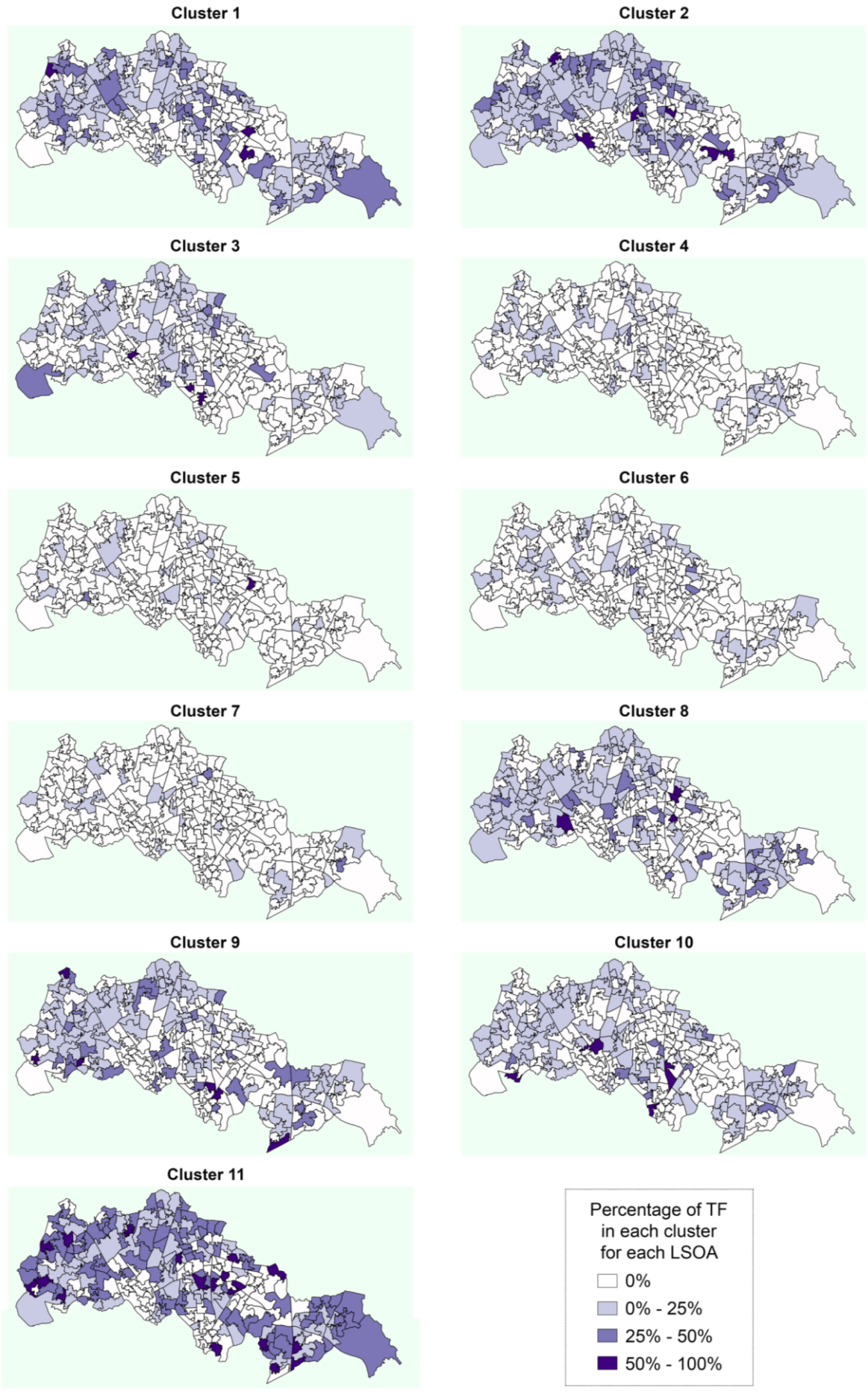


Figure 52: TF living in each LSOA as a percentage of the total TF living there, by cluster (utilising ECC data)

Whilst Figures 50-52 provide visualisations of the location and density of the TF in relation to their cluster assignment, Table 21 displays a selection of the demographic data (from Census 2011) aggregated by cluster. This utilised the OA level data, as it provided more fine-grained detail than the LSOA level data. The attributes had different units: ethnic group, place of birth and qualifications were compiled as the percentage of people with these characteristics in an area; economic activity applies to the household reference person for each household; and household deprivation and tenure apply on a household level. In each column, the maximum and minimum values for that characteristic are highlighted in bold. For instance, for the population density attribute, families from cluster 3 lived in areas with the highest average population density (83.3 persons per hectare) and families from cluster 10 lived in areas with the lowest average population density (67.4 persons per hectare).

Figure 53 plots the data from Table 21 in a parallel points plot, with a line for each cluster (each assigned a unique colour to differentiate), this was plotted in order to provide a more interpretable comparison of the clusters (so much information in a table can be difficult to comprehend). Figure 53 also contains the long-term health attribute (percentage of people who felt their long-term health was limited either a little or a lot), which was not contained in the table due to space constraints.

Table 21: Aggregated demographic data by cluster assignment with interesting characteristics highlighted in bold (utilising ECC and Census 2011 data)

Cluster	**Population Density:** persons per hectare	**Economic Activity:** Percentage economically active	**Ethnic Group:** Percentage White	**Place of birth:** Percentage born in UK	**Qualifications:** Percentage with no qualifications	**Household Deprivation:** Percentage deprived in at least one dimension	**Household Tenure:** Percentage that own home	**Household Tenure:** Percentage renting social housing
1	68.8	60.0	**69.6**	**77.8**	34.7	73.9	31.7	48.5
2	78.1	61.5	67.3	76.0	33.0	72.7	31.5	44.3
3	**83.3**	**62.7**	**61.5**	**72.8**	31.2	71.9	32.1	42.7
4	75.4	**57.7**	63.4	74.9	**35.0**	**75.9**	**27.6**	**53.1**
5	70.2	60.0	65.2	73.7	**30.2**	**71.4**	29.0	48.6
6	76.6	61.3	64.6	75.1	33.3	72.3	30.6	47.2
7	75.2	61.4	66.2	76.2	33.5	72.5	29.2	50.9
8	73.7	61.2	67.9	77.0	33.6	73.8	31.0	47.6
9	75.5	62.2	64.7	74.1	32.0	72.5	**33.0**	**41.1**
10	**67.4**	60.4	68.7	77.4	33.7	73.9	30.3	49.2
11	68.9	60.2	68.7	77.5	34.2	74.2	30.4	49.7

It is important to consider that since clusters 4 to 7 were relatively small ($n < 70$), the statistics pertaining to them may be less reliable than for the other larger clusters. However, given these caveats, perhaps the most notable characteristics apply to cluster 4,

which was one of the smaller clusters (n=61) and whose main characteristic was that each family had at least one member who was NEET (not in education, employment or training). Families from cluster 4 lived in areas with the highest levels of household deprivation of all the clusters, highest levels of people with no qualifications, highest levels of people who felt their long-term health was limited, and also the highest levels of households living in social housing. Cluster 4's families also lived in areas that had the lowest levels of economic activity of all the clusters.

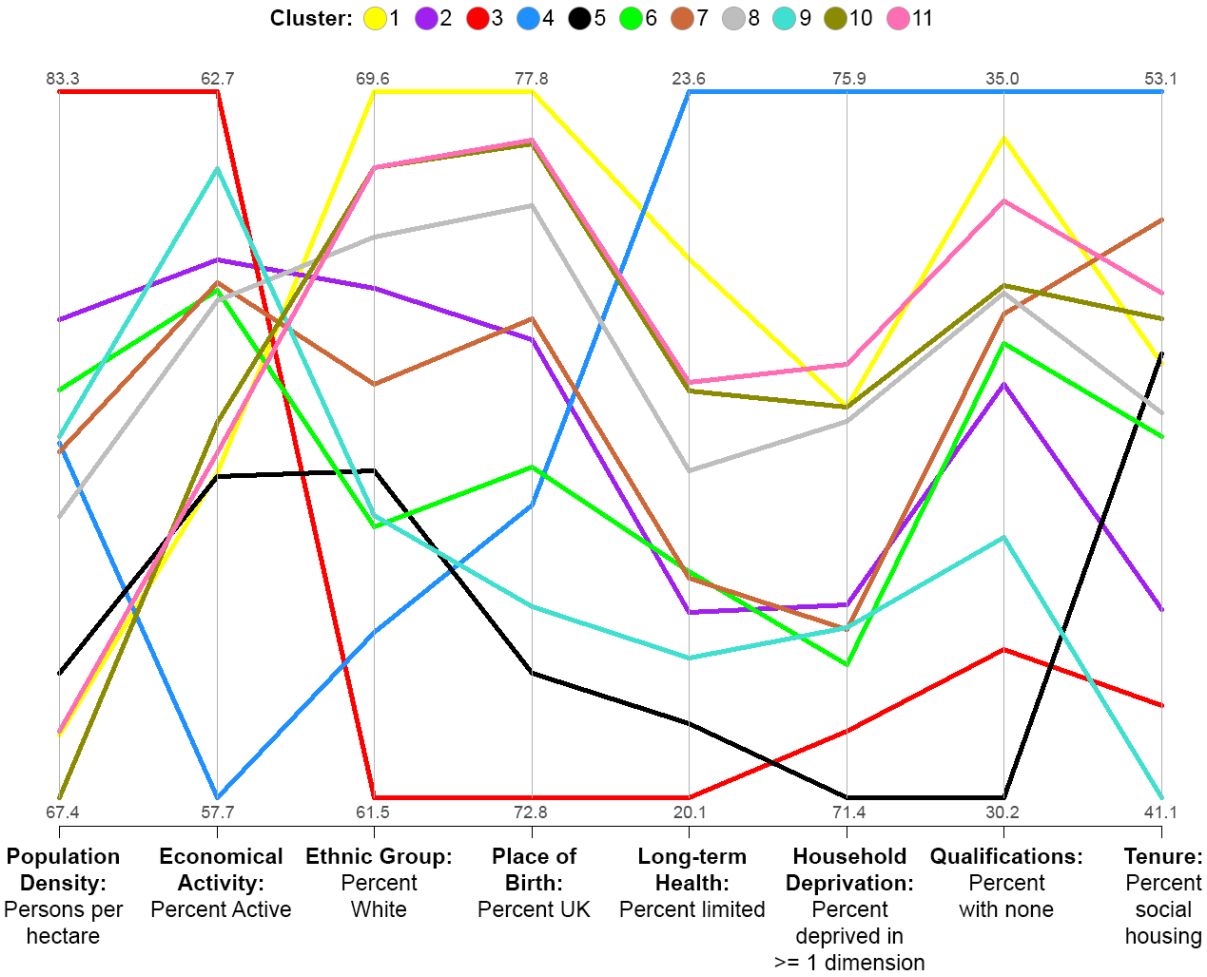


Figure 53: Parallel points plot of place-based data (Census 2011 data linked to the Output Area that each TF lived in) aggregated by cluster assignment

Cluster 3, which contained families who all had Looked After Children events, also had notable statistics. In comparison to the other clusters, families in cluster 3 tended to live in areas that had higher population density, and the lowest percentages of people with 'white' ethnicity, and people who were born in the UK. Families from cluster 3 also lived in areas that had the highest levels of economic activity, and comparatively high levels of home ownership, but lower levels of households living in social housing and the lowest levels of people who felt their long-term health was limited. In contrast, families from

cluster 1 lived in areas that had the highest levels (over all 11 clusters) of people of 'white' ethnicity and people born in the UK. The average population density for the areas that families from cluster 1 lived in was lower than for families in all but two of the other clusters, and families from cluster 1 lived in areas that also had higher average levels of people with no qualifications.

Figure 54 plots heatmaps of the locations of families in clusters 1 and 3, together with a combined plot in order to highlight the differences in location. In the third plot, cluster 3 was layered over cluster 1, therefore it was rendered partially transparent in order to be able to visualise cluster 1 underneath; where the colours are muddy is where both clusters had a high concentration of families (this is most evident on the right side, vertical middle of the plot). Whilst it should be considered that cluster 1 contained more families than cluster 3 (291 compared to 115), and so might have a greater spread, there were subtle differences in location hotspots for each cluster. Since there was no identifying geography included in the maps (no underlying map), a grid was included on each map in order to be able to compare the locations accurately (the gridlines follow the same geographical positions on each map).

Whilst there were differences between clusters 1 and 3, it was notable, in contrast, that families from clusters 10 and 11 lived in areas with very similar characteristics; the two lines for these clusters in the parallel points plot follow almost exactly the same course. Cluster 1 also follows a similar course, although deviates slightly at certain points.

Table 21 also highlighted that families from cluster 9 (which contained families who had only school absence and CIN events), lived in areas that had the highest levels of home ownership and conversely had the lowest levels of households living in social housing of all the clusters. They also lived in areas that had comparatively high levels of people who were economically active. Families from cluster 5, which was the smallest cluster (n=21) and contained families who all had at least one adult who had committed a criminal offence, lived in areas with the least household deprivation of all clusters, and the lowest percentage of individuals with no qualifications.

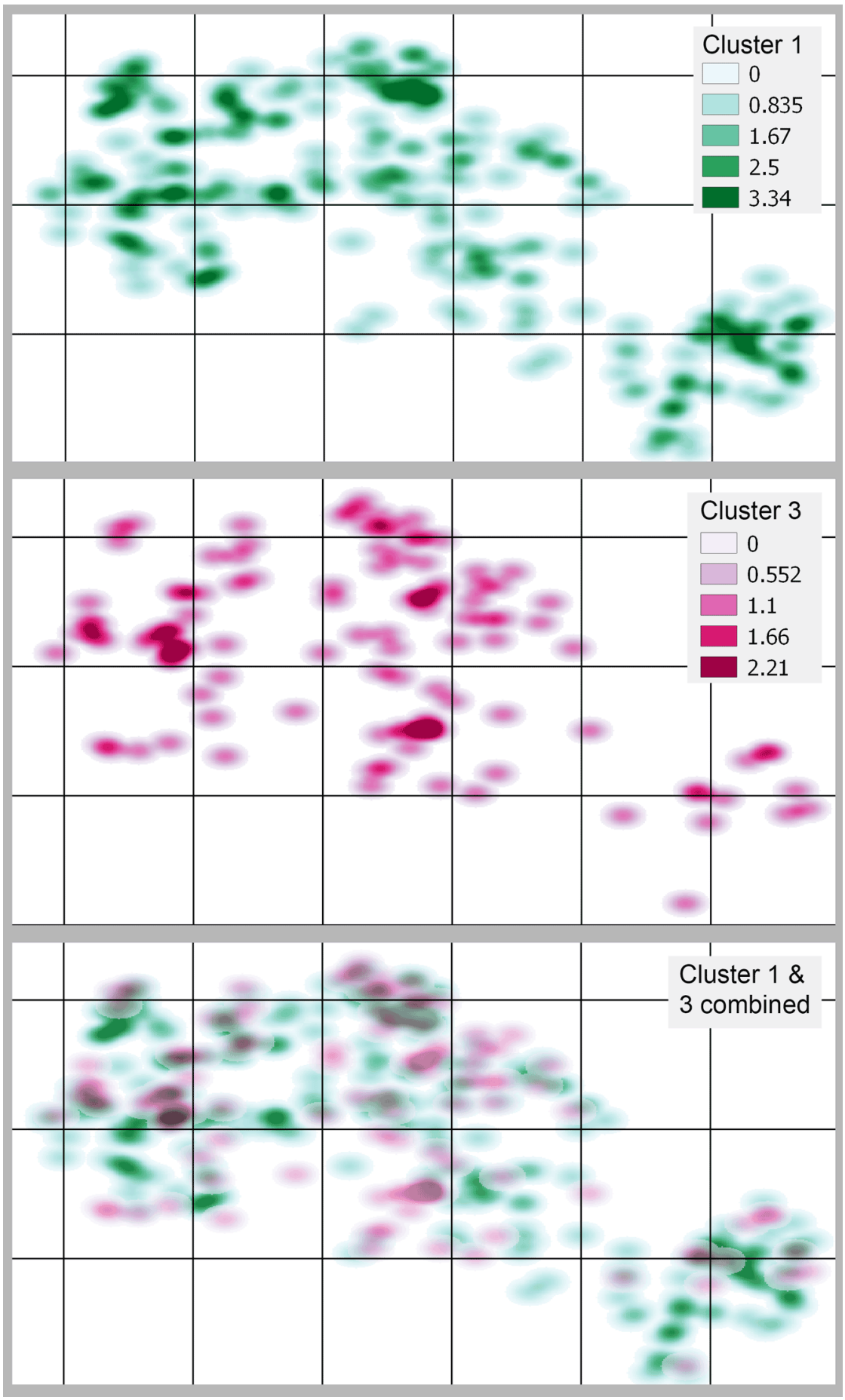


Figure 54: Heatmap comparison of the geographical locations of families in clusters 1 and 3

Overall, this section has highlighted that there were subtle differences in the areas that families from each cluster lived in. Whilst there was no magical division of families into unique areas by cluster assignment (for example, families from cluster 1 did not all live in one small area, and families from cluster 2 in another separate small area, etc.), the analysis did indicate subtle differences in location by cluster. For the most part, families from each cluster were spread throughout the city, however the various plots highlighted that there were differences in location and concentration of TF by cluster, and the statistics compiled for the areas that the families lived in also indicated differences.

In order to gain a better understanding of the particular factors that might be important to these differences (or that are predictors of cluster assignment), the following section utilises machine learning methods as a way to identify important 'place-based' attributes.

6.3.4.1 Predictive Modelling to determine whether place-based data might be considered predictors of cluster assignment

Machine learning models (decision trees, random forests and generalized boosted models) were built in order to determine whether the place-based data might be utilised to predict cluster assignment. This was performed in order to determine which of the place-based attributes were the most important with regards to the clusters. The focus was not prediction, as there was no real necessity to be able to predict cluster assignment from the place-based data. However, these methods each rank the predictor attributes in terms of importance to the target, and this can provide useful insight as to whether any of the attributes are important. Decision tree learning, random forests and generalised boosted models were chosen specifically because they each produce an importance measure for the predictor attributes.

For comparison purposes (to a more traditional regression method) multinomial logistic regression was also performed; this method was chosen as it can deal with an unordered categorical target attribute with more than two levels. However, it should be considered that the model was misspecified as many of the predictors were correlated (as highlighted previously in Figure 10), and the target attribute (cluster assignment) had two levels that were particularly small (clusters 5 and 7). Small levels such as this can lead to unreliable results (Boslaugh, 2013); one solution would be to merge the levels, however it made little sense to merge the clusters.

The cluster assignment (numbered 1 to 11) was the target attribute. The various place-based attributes (taken from the 2011 Census data linked to OA, and 2011 Police data linked to LSOA) were the predictor attributes, these are detailed in Appendix A. The data was randomly split into a training (n = 1509) and testing (n = 646) dataset, with a 70:30 split, and this was applied proportionately to the target attribute. Each model was built upon the training dataset and then a final model evaluation performed using the test dataset.

The baseline accuracy (on the test dataset) for predicting cluster assignments was 28.02%, therefore any model that performed better than this might be considered useful. Baseline accuracy was derived by considering the simplest model possible; i.e., the accuracy if the model simply predicted the largest cluster (in this case, cluster 11).

Various attribute combinations and different weightings were experimented with, however none of the models could improve significantly upon the baseline accuracy of 28.02%. Table 22 details the best accuracy on the test dataset of each of the models. The decision tree had the highest accuracy overall, which was a tiny improvement upon the baseline. The random forest model had the worst performance, with the accuracy being nearly 9% lower than the baseline. The multinomial logistic regression model had accuracy that compared favourably with the boosted model. However, it would be fair to say that none of the models performed better than simply guessing.

Table 22: Accuracy on test dataset for each of the models predicting cluster membership using place-based attributes

	Method:				
	Baseline Accuracy	**Decision Tree**	**Random Forest**	**Generalized Boosted Model**	**Multinomial Logistic Regression**
Accuracy on test dataset	28.02%	28.17%	19.06%	26.78%	26.32%

It was noted that the random forest models performed in a different manner to the other models. Whilst the other models essentially just predicted the largest cluster (cluster 11, and therefore just about matched the baseline accuracy), the random forest also attempted to predict some of the other cluster assignments (albeit with low accuracy), and so in some ways might have detected more of an underlying pattern. However, none of the models performed well enough to consider them useful in terms of prediction. In terms of variable importance, all machine learning models chose similar attributes as the

most important; Table 23 lists the top five predictors for each model. Appendix A contains the full list of important predictors for each model, together with the model parameters and results. Overall, attributes relating to housing tenure, health, economic activity, household size and qualifications were deemed most important, and despite the poor performance of the models, this would appear to reinforce the patterns already noticed within the place-based data in the previous sections.

Table 23: Most important 'place-based' attributes for each model to predict cluster assignment

Decision Tree	Random Forest	Boosted model
Tenure – percentage households who private rent	Long-term health – percentage people that are limited	Tenure – percentage households who private rent
Long-term health – percentage people that are limited	Economic activity – percentage people who are active	Long-term health –percentage people that are limited
General health – percentage people with bad or very bad	Place of birth – percentage people born in UK	Household size – percentage single person households
Qualifications – percentage people with none	Household size – percentage single person households	Economic activity –percentage people who are active
Tenure – percentage of household who social rent	Qualifications – percentage people with none	General health – percentage people with bad or very bad

The multinomial logistic regression model produced a complex model. Ideally a smaller set of predictors would be utilised, and a dataset with fewer small groups. However, this method was chosen for direct comparison to the machine learning methods, and so exactly the same data was utilised for all models. As already noted, the model was misspecified and whilst the model did indicate significant attributes for some of the clusters, the results may not be reliable (they are listed fully in Appendix A). However, it did compare favourably with the other methods in terms of predictive accuracy on the test dataset.

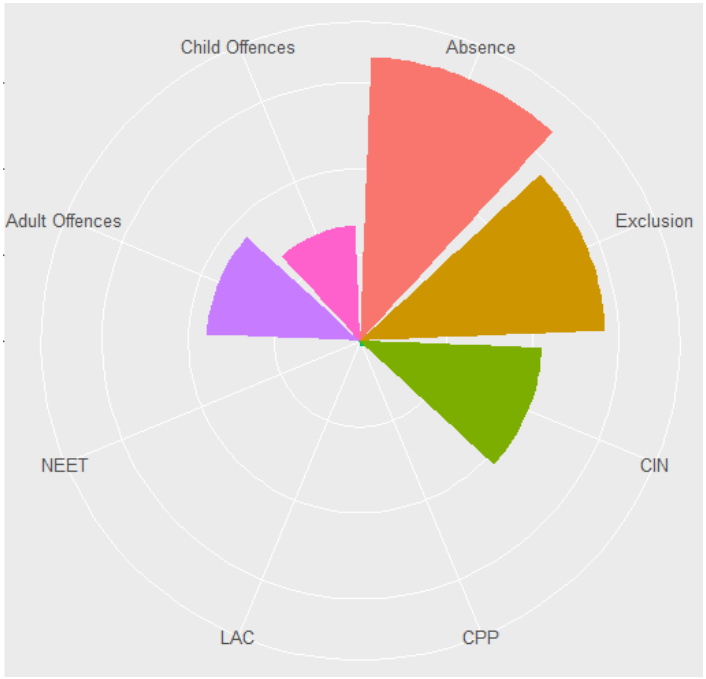
A second logistic regression model was built in order to determine if utilising the machine learning methods as a form of feature selection (in order to select a smaller number of predictors) might improve the model performance. Only the top five predictors identified as most important (listed in Table 23) by each of the machine learning models were included; since the lists were very similar, this consisted of 8 predictors. This model produced accuracy on the test dataset of 27.86%, which was an improvement of 1.5% on the original regression model, but still below the baseline accuracy. Full details are contained in Appendix A.

Overall, this section has shown that it was not possible to predict the cluster assignments from the place-based data alone. This was perhaps a likely outcome as the place-based data was somewhat distant from the family itself - it could not say anything specifically about an individual family, only that the family lived in an area with particular characteristics (for example, low economic activity). However, prediction was not the purpose; identifying the attributes that were predictors of the cluster assignment was. Whilst it should be considered that the model performance was generally poor (no improvement upon guessing), the identified predictors were similar no matter the method, suggesting they may have picked up some underlying pattern, and this may go some way toward confirming the findings from the previous sections; that place-based characteristics such as housing tenure, health and economic activity had some importance to the clusters.

6.4 SUMMARY

There follows a final brief summary of the clusters, drawing together the information from the previous sections and including the place-based information. For clusters 1 to 7 a plot is included that shows the percentage of families in that cluster with each event in order to provide a visual reminder of the cluster characteristics.

6.4.1.1 Cluster 1: School exclusion and criminal offences (n = 291)



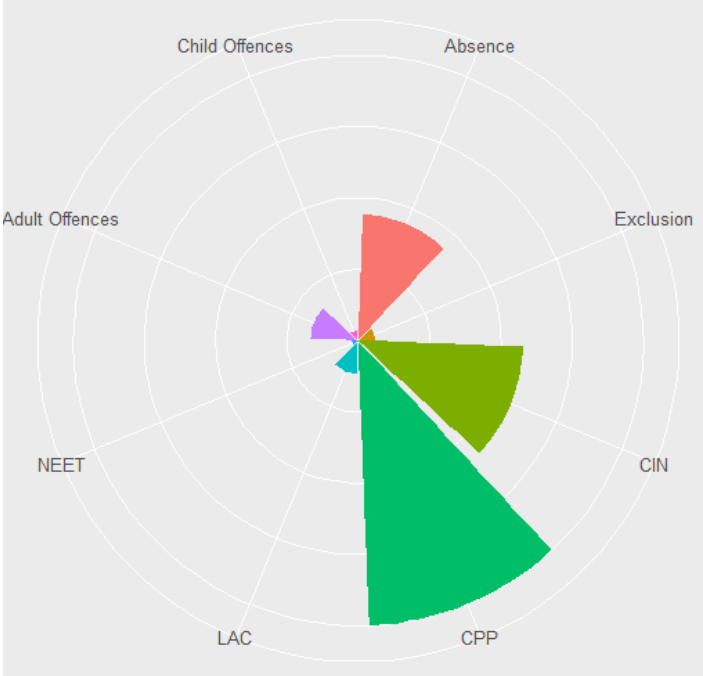
Crime and problems with schooling were the main feature of this cluster. Families had high levels of criminal offences committed by adults and criminal offences committed by children, although they tended to have one or the other, but not both. The families had the highest levels (comparably) of school exclusion and most of those families with school exclusion also had school absence, although the levels of absence were relatively low. It was notable that

proportionately more children (than any other cluster) attended schools that were classed by OFSTED as requiring improvement or inadequate. In terms of child safeguarding, there were average levels of CIN events, but very low levels of more serious events (CPP and LAC). No family had any NEET members.

Families from cluster 1 lived in areas that had the highest levels (over all 11 clusters) of people of 'white' ethnicity and people born in the UK; and lower average population density. There were also higher levels of people with no qualifications.

This cluster had a fairly low average silhouette value (0.22) and might be considered only loosely cohesive; it was likely that some of these families may have been better suited to another cluster.

6.4.1.2 Cluster 2: Child Protection (n = 335)



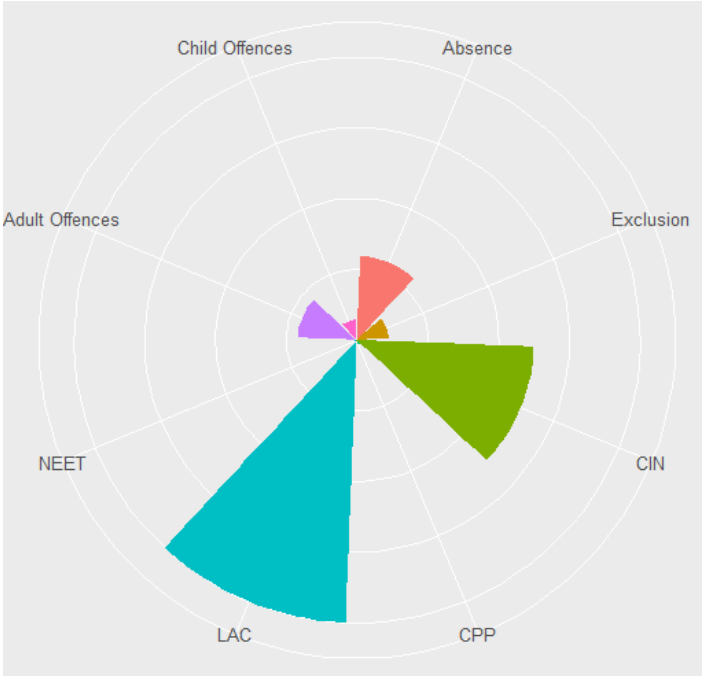
Child safeguarding was the main feature of this cluster. All families had Child Protection Plans, and there were proportionately higher levels of CIN events and children who had been taken into care (LAC) in comparison with other clusters. The families tended to have younger children with over three quarters aged 11 or under. Perhaps because of their younger ages, there was very little school exclusion and criminal offences committed by children. However, just under half of families had school absence, but the levels of absence were low.

There was no remarkable aspect where the place-based data was considered, although families from cluster 2 lived in areas that on average had higher population density, and more economically active people than for most of the other clusters.

Cluster 2 was a diverse cluster, with 44% of families having 3 or more different types of events occur in the year prior to intervention. This more complex mixture of events was reflected in that cluster 2 had the fewest (proportionately) families receiving AO

treatment; AO was aimed at families whose needs were at risk of becoming complex, therefore it would seem that many of the family's needs were already judged to be complex. Families in cluster 2 had proportionately more planned endings to their intervention treatment than all but one of the other clusters. The average silhouette width of 0.44 was acceptable and it is likely that most families did belong in this cluster.

6.4.1.3 Cluster 3: Looked after Children (n = 115)

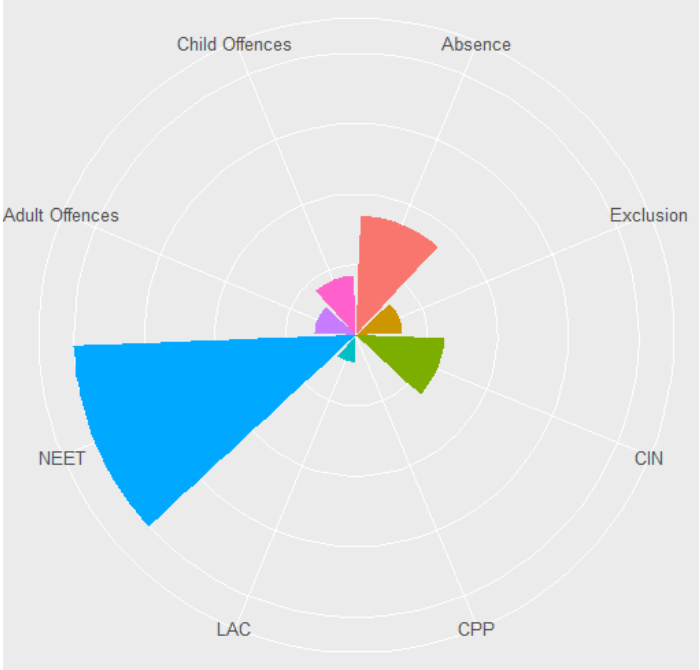


The main feature of this cluster was that all families had a child (or children) taken into care at some point (LAC events) in the year prior to intervention. A high proportion of families also had CIN events, but there were almost no CPP events. The other main feature was criminal offences committed by adults; just over a fifth of families had them, which was comparably high. There were low levels of school absence and exclusion. Just under a quarter of children attended schools that were classed as 'Outstanding' by OFSTED, higher than any other cluster.

More families in this cluster had address changes than any other cluster, however this might be at least partially explained by address changes related to children moving in and out of social care. In comparison to the other clusters, families in cluster 3 tended to live in areas that had higher population density, and higher levels of economic activity. Conversely, the areas they lived in had lower percentages of people with 'white' ethnicity, people who were born in the UK, and people who considered their long-term health to be 'limited'.

Cluster 3 had the highest silhouette value (0.56) of all the clusters and so might be thought of as the most cohesive cluster. It is likely that almost all families belonged in this cluster. Half of the families received FF (Families First) treatment, which is aimed at keeping families together (where safe), and overall there were proportionately more planned endings for treatment than any other cluster.

6.4.1.4 Cluster 4: NEET (n = 61)

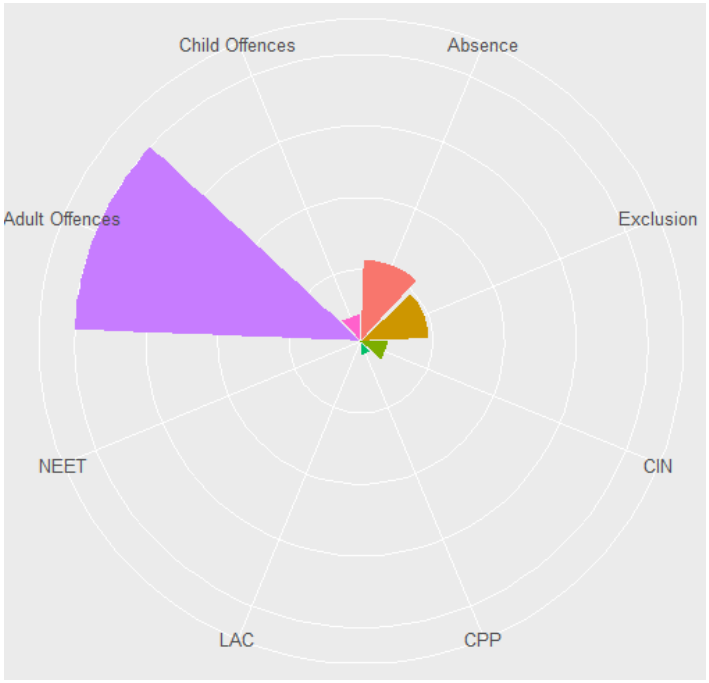


The main feature of cluster 4 was that all families had at least one member who was not in employment, education or training (NEET). This cluster contained a particularly high proportion of children in their late teens (aged 16-17), and young adults (aged 18-20). Perhaps because of the older ages of the children, there were low levels of child safeguarding events (although LAC events were slightly above average, at 11%). School absence and exclusion levels were average, however a fifth of families had children who had committed criminal offences, which was comparatively high.

In comparison to the other clusters, families from cluster 4 tended to live in areas that had higher levels of household deprivation, people with no qualifications, households living in social housing, people whose long-term health was 'limited', and people who were economically inactive. Cluster 4 had fewest (proportionately) children attending schools judged as 'outstanding' by OFSTED. Cluster 4 had an acceptable silhouette value (0.44) and appears fairly cohesive.

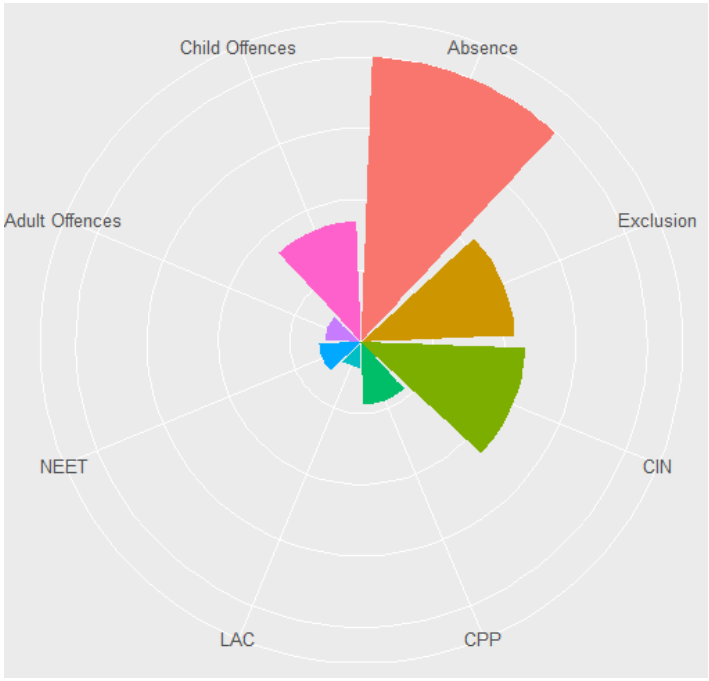
6.4.1.5 Cluster 5: Adult criminal offences (n = 21)

The main feature of cluster 5 was that all families had at least one adult who had committed criminal offences. These were at a high level, with a mean of 4 offences per family. A third of families had events which were classed as domestic abuse; this was more than any other cluster. Half of the families had no children (aged under 18), perhaps because of this there were very few child safeguarding (CIN, CPP, LAC) events. However, school exclusion levels were high; of those families with children, half had children who had been excluded.



This cluster was not very diverse, with almost two thirds of families having just criminal offences committed by adults and no other type of events. It was also the smallest of all the clusters, with only 21 families. Cluster 5 had a good silhouette value of 0.46, and it is likely that all families were probably best suited to this cluster. The families lived in areas with lower levels of household deprivation and people with no qualifications, however given the small size of the cluster (n=21), a higher sample size would have been desirable to ensure these statistics were reliable.

6.4.1.6 Cluster 6: High levels of school absence (n = 54)



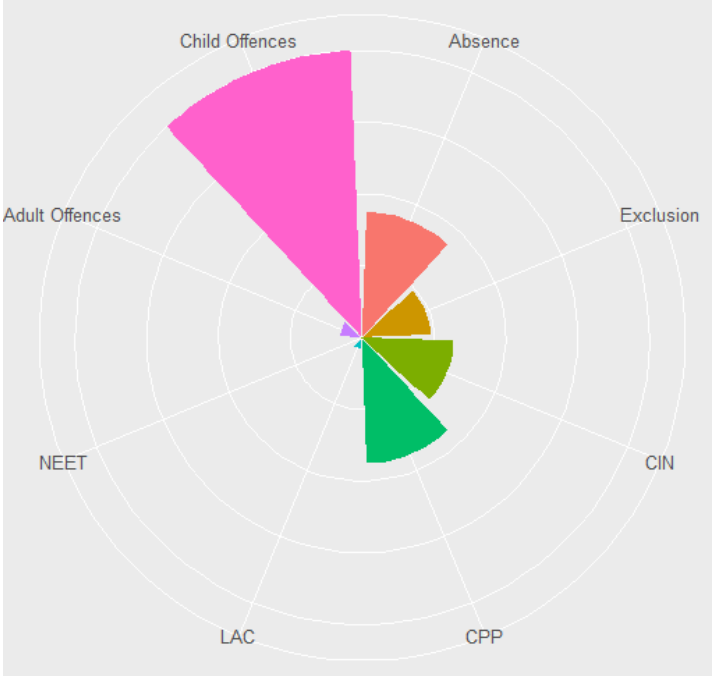
Whilst the most noticeable feature was that all families had high levels of unauthorised school absence (39% per family on average), another aspect of cluster 6 was that the occurrence of all events was greater than average (i.e. for all families). The families generally had a complex mixture of events, with three quarters having 3 or more different types of events. Alongside school absence, school exclusion and criminal offences committed by children were also high. Child safeguarding events were also comparatively high.

There was no particular place-based aspect of the data that stood out for cluster 6, although its families tended to live in areas that had comparatively lower levels of

household deprivation. However, the percentage of children from cluster 6 attending schools judged as 'good' or 'outstanding' by OFSTED was lower than for all but one of the other clusters.

The silhouette value of this cluster (0.14) was the lowest of all clusters, and the low value implies that this cluster was not very cohesive and that some of its families may have been better suited to other clusters. A high proportion (more than any other cluster) of families received FIP (Family Intervention Project) treatment; since this was aimed at families with the most complex needs, it reflects the complexity of needs for families in this cluster.

6.4.1.7 Cluster 7: Child criminal offences (n = 25)



The main feature was that all families had criminal offences that were committed by children, and these were at a high level (with a mean of 4 per family). However, few families had criminal offences that were committed by adults. The other notable feature was a comparatively high proportion of families with Child Protection Plans (just under half), and school exclusion (just under a quarter). Those families with CPPs tended to also have school absence, although absence levels were fairly low. The other child safeguarding events (CIN and LAC) were comparatively low. A fifth of families had events classed as domestic abuse; only one cluster had a higher proportion than this.

Families from cluster 7 lived in areas that on average had higher levels of households living in social housing. The percentage of children from cluster 7 attending schools judged as 'good' or 'outstanding' by OFSTED was higher than for all other clusters. However, this was a particularly small cluster, with only 25 families, and a higher sample size would have been desirable to ensure the reliability of these statistics. The average

silhouette value of the cluster was fairly low (0.20), which indicates a possible lack of cohesion.

6.4.1.8 Cluster 8: School absence only (n = 223)

The main feature was that all families had school absence and no other events. The average unauthorised absence per family was 6.4% of available sessions; three of the other clusters had higher averages. There was no remarkable aspect of the place-based data related to cluster 8, and the percentage of children attending schools judged as 'good' or 'outstanding' by OFSTED was average in comparison to the other clusters. One notable aspect was that 13% of families had no adult attached to them, which was the highest proportion across all clusters. A high proportion of families received CFPT (Complex Families Parenting Team) treatment; since this is aimed at families with a range of complex needs, it may suggest that at least some of these families had other needs that were not captured by the available data.

6.4.1.9 Cluster 9: Children In Need events only (n = 243)

The main feature was that all families had CIN events and no other events. A large proportion of children in this cluster were aged under 7; a third of all children were too young to attend school. Of those that did attend school, one fifth attended schools judged by OFSTED to be 'outstanding', this was comparatively high.

Families from cluster 9 lived in areas that on average had comparatively high levels of people who were economically active, and where home ownership was higher. As noted with cluster 8, a high proportion of families received CFPT (Complex Families Parenting Team) treatment; since this is aimed at families with a range of complex needs, it may suggest that at least some of these families had other needs that were not captured by the available data.

6.4.1.10 Cluster 10: Absence and CIN (n = 182)

The main feature was that all families had school absence, and at least one CIN event, but no other events in the year prior to intervention. The average percentage of unauthorised absence was 10.6% per family, which was relatively high.

Families from cluster 10 lived in areas that on average had lower population density, and comparatively high levels of people who were born in the UK and in the White ethnic group. A comparatively large proportion of families received either CFPT or FIP

intervention treatment, which are both aimed at the families with more complex needs, suggesting that at least some of these families may have had other problems that were not reflected in the available data.

6.4.1.11 Cluster 11: No events (n = 605)

All families had none of the events in the year prior to intervention. The types of families in this cluster were somewhat different to the other clusters: 41% of families consisted of a single person, this was a far higher percentage than for any other cluster (cluster 5 had 14%); and half of all families had no children attached to them. One possible reason for so many single person families (which are generally not typical of the TF programme), may be errors in the available data, i.e. there may have been missing links between individuals in some cases.

In comparison to the other clusters, families from cluster 11 tended to live in areas with higher levels of household deprivation, people born in the UK and those belonging to the White ethnic group; and with lower average population density. The percentage of children from this cluster attending schools judged as 'good' or 'outstanding' by OFSTED was high in comparison to the other clusters. A comparatively large proportion of families received AO (Assertive Outreach) treatment, which is aimed at those families who are at risk of developing complex needs. This may indicate that although it is not clear from the available data what the family's problems were, there was concern that their problems could escalate.

6.5 DISCUSSION

The eleven clusters found in this analysis each had unique characteristics, from those with a diverse range of events, to those with few or none. Some were very cohesive, for example, cluster 8 whose families all had only school absence; whereas some were more loosely joined together, for example cluster 6, with the lowest silhouette value. The discovery of these clusters means that future analysis might be carried out on the cluster-level, rather than the global (all families) level, and may mean that separate models (such as decision trees or regression analysis, etc.) might perform better than one overall model, since the groups should be more heterogeneous. Analysis in the following chapter explores this idea.

It should be considered that the clusters found in the analysis were a product of the particular data that was available. The data-driven clustering model utilised only the attributes that were considered complete (that is, did not have known missing data) for the year before the start of a family's first intervention. Therefore, the cluster analysis had a focus on child safeguarding (CIN, CPP, LAC), crime (committed by children or adults), and education (school absence, exclusion and NEETs). Had different data been available, the focus would have been different, and hence the clusters may have been different too.

The analysis indicated that there was a large proportion of families who did not have a diverse mixture of events. Just over a quarter (28%) of families had none of the listed events in the year prior to first intervention (cluster 11), whilst 30% had only one event, and 24% had two different events. Only 19% of families had 3 or more different events (the majority of these being 3 events) and so might be considered to have had a wider range of complex needs. Overall, the fact that most of the families did not have a wide range of events meant that the families generally fell into clusters that were relatively simple to describe (for example, no events, or just school absence), or else had one main feature (such as all families with CPPs, or LAC events). This meant that the decision tree analysis was able to provide a set of relatively simple rules to accurately describe the cluster assignments. The decision tree identified Child Protection (CPP) as the most important attribute for predicting cluster assignment; perhaps the main reason for this was that cluster 2 was the largest cluster (of the 7 clusters produced from the cluster analysis; it contained 335 families) and its main feature was that all families had Child Protection Plans, therefore it would make sense that the first split in the tree would attempt to split the largest group of families off. The second most important attribute was families having looked after children (LAC); together with the CPP attribute, this indicated that child safeguarding was a particularly important feature where considering clusters 1 to 7.

However, whilst the decision tree could provide an indication of 'important' attributes, it should be considered that clusters 8, 9, 10 and 11 were not included in the decision tree analysis, as their rules were so simple (that is: just school absence; just CIN events; just absence and CIN events; and no events). Yet they accounted for a large portion of families (58%), and so, if the most important attributes were to be considered for these

clusters, they would be school absence and CIN events. However, it should be considered that perhaps the most important feature overall was that which characterised the largest cluster (11) and hence accounted for the most families; that is, having no events (no child safeguarding, crime or school concerns in the year prior to intervention).

Whilst the families in cluster 11 had none of the eight events considered, it is unlikely that they truly had no events in the year prior to intervention. As already considered, one possible reason for the lack of events may be missing data, but this is unlikely to account for all families in cluster 11. The criteria for joining the TF programme indicated that families should generally have multiple problems, and the large groups of families within the ECC data with no, or few events, indicates that some of these families may well have had other problems that were simply not captured by the available data.

Research into TF generally has found that many have very complex needs, far more than could be represented by the available data in this study. In interviews with 20 families Boddy et al. (2016:285) found that families consistently had health problems; they had 'unrecognised, unmet, and/or poorly managed health needs, relating to key aspects of basic health and significant and chronic physical and mental health problems'. Similarly Wenham (2017) interviewed ten young people involved in the programme and found that their problems included bereavement, financial hardship, child abuse and domestic violence. Shildrick et al. (2016) also make the point that in the case of some families the sheer number and complexity of their problems was high.

Another indication of the types of events that some of the families may have is contained in the National Evaluation report (Department for Communities and Local Government, 2017). This states that, overall for TF in England in the year before intervention, 40% of TF had a family member with a mental health issue, 34% had police called out to their home, 25% had a member involved in domestic abuse, 12% had an individual dependent on drugs or alcohol, and 10% had Anti-Social Behaviour incidents. Since there was no, or incomplete, data in the ECC database regarding these problems, it may be that similar proportions of ECC families also had these problems. This may go some way towards explaining at least some of the families with no, or few events; however, it could also mean that those families with some events had an even more complex mixture of events.

Overall, the analysis indicated that TF tend to live in areas with higher percentages of lone-parents, higher levels of deprivation, lower educational levels, poor health, less

economic activity and higher levels of social housing, as might have been expected. These areas were loosely focussed in the North-West and Eastern areas of the city. Analysis utilising the different clusters did indicate subtle differences in where families lived. Perhaps most notable was that, in general, families belonging to clusters whose main feature was child safeguarding (clusters 2, 3 and 9) tended to live in areas with (comparatively) higher levels of economic activity and higher population density, and lower levels of social housing. Families from Clusters 1 (crime and school exclusion), 10 (school absence and CIN), and 11 (no events) tended to live in areas with (comparatively) higher proportions of people born in the UK and in the White ethnic group. Yet, unlike clusters 2, 3 and 9, there was less to tie these three clusters together, aside from the lack of the more serious child safeguarding problems (CPP and LAC).

Families from Cluster 4, which consisted of families with at least one NEET member tended to live in areas with (comparatively) high levels of household deprivation, people who had limited health and people with no qualifications, together with lower economic activity. This might pose the question of whether NEET status may be related to where an individual lives; research indicates that predictors of becoming NEET include growing up in areas with poverty, lower socioeconomic status and lacking good schools (Sadler et al., 2015), and the place-based analysis also appears to indicate this (but would require further research to fully consider this question).

Another question that was suggested by the analysis was whether having CIN events meant that unauthorised school absence levels were higher for some families. This came from considering clusters 8 (just school absence) and 10 (just school absence and CIN events). Families in cluster 8 had an average of 6.4% school absence, whereas families in cluster 10 had 10.6% on average. On the surface, cluster 10 only differed from cluster 8 in that the families also had at least one CIN event (although of course there may have been other underlying factors), therefore it might be considered that the addition of CIN events in this cluster could have been a factor in the increased level of unauthorised school absence.

6.6 Conclusion

The extensive practical work in this chapter explored the data pertaining to the Troubled Families programme of an English City. It considered the characteristics of the families

and whether there existed different groups, or clusters, of families in the data. It also considered where the families lived and whether there were any demographic patterns, and if where a family lived played a part in deciding which cluster they belonged to.

The data-driven cluster analysis identified eleven clusters of families, all with different characteristics. Whilst some were more cohesive than others, there was no doubting that there existed unique groups of families within the data. The focus of the clusters centred on child safeguarding, crime and education. These particular attributes were chosen as they were considered to be complete (that is, had no known missing data). Data for the year prior to a family's first intervention was utilised in order to represent the events that led to a family's introduction to the TF programme. Each of the clusters had particular characteristics that were unique to them, for instance, cluster 2 contained families who all had Child Protection Plans, whereas cluster 11 contained families who all had none of the events in the year prior to the start of intervention.

The available data indicated that, for at least some of the TF, their needs were perhaps not as complex as might have been expected when considering the Government criteria for entry into the TF programme. A large proportion of families had no, or few, events in the year prior to their introduction to the TF programme. This appeared to indicate a lack of diversity of events for these families, as generally they should have had at least three different types of events to qualify. However, there were various problems with the data, and some key attributes that contributed to the definition of a Troubled Family were not available or missing (such as, incomplete benefits and anti-social behaviour data, and the absence of any data pertaining to health). Therefore, it was not possible to determine whether at least some of the families who appeared to have no (or few) issues might have had other issues were more data available, and in such case would have met the Government guidelines of belonging to the TF programme.

In general, this study has highlighted some of the problems with the available ECC data, such as: missing data (particularly historical benefit data); duplicate people; and likely missing links between family members. The ECC found this useful as it helped them to identify problems in their data and consider ways to fix them.

This case study has shown that decision tree learning can be employed to derive rules for cluster assignments; the tree visualisation and accompanying rules accurately assigned families to their particular cluster in a simple manner. The fact that the tree was easily

understandable and quite small indicated that, despite the seemingly complex nature of the various clusters (as evidenced by the long cluster descriptions in section 6.3.2), they could be described by a set of relatively simple rules. This work highlighted that decision tree learning can be utilised in order to provide clarity to complicated datasets. It also produced a re-usable model which could be utilised (if required) to assign future families to the appropriate cluster.

Data visualisation methods such as t-SNE and Nightingale plots were utilised in order to present clustering results in a way that might aid better understanding of the data and clusters. t-SNE was able to represent higher-dimensional data on a two-dimensional plot in a way that made sense, but also highlighted that some of the clusters were not neatly separated into distinct blocks; it indicated overlaps, and that some of the clusters were not as cohesive as others.

The overall place-based analysis of the families in relation to where they lived at the start of intervention indicated that the families tended to live in areas with higher percentages of lone-parents, higher levels of deprivation, lower educational levels, poor health, less economic activity and higher levels of social housing. To some degree, results such as this might be expected. It also highlighted that there were two areas of the city that had higher proportions of TF; the North-Western corner and the East. Further detail was provided by the place-based analysis of the separate clusters; this indicated that families from different clusters were concentrated in subtly different areas of the city and that the characteristics of these areas could vary quite widely by cluster. For instance, families whose main problem was child safeguarding (CPP, LAC, CIN, contained in clusters 2, 3 and 9) appeared to live in areas with higher levels of economic activity and higher population density. And families with NEET members (cluster 4) appeared to live in areas with higher levels of deprivation and lower economic activity. It would seem therefore that the cluster assignment did indicate patterns and relationships within the place-based data; these patterns would warrant more detailed further research.

It might be considered that the most important features of the eleven clusters were related to child safeguarding, school absence, and the absence of any problems (no events). However, it was highlighted that to really understand the underlying context within the data, other attributes should be considered (such as those pertaining to health, anti-social behaviour, receipt of benefits, etc.). If data such as this were to become

available in the future, an opportunity for further work would be to analyse the existing clusters with reference to this, and also to consider whether different clusters might exist.

Overall this chapter has highlighted that there existed different groups of families within the data and that identifying and analysing each of the groups provided a deeper understanding and far more context than could be achieved by simply performing a 'global' analysis of the entire group of families as a whole. The information gained from studying these clusters of families might be used to inform those working with the families; the ECC found the study informative and have begun to adopt methods such as clustering in their own analysis.

More broadly, the identification of different clusters (or types or groups) of families may mean that it is possible to identify where particular treatments or methods might work better (or equally be less likely to succeed) for the different clusters. A deeper understanding of the types of families may mean that the particular treatment received, or how it is administered, could correspond to the cluster a family is in; this might lead to more effective treatment.

The work in this chapter created a foundation for the following chapter which explores the families and their cluster assignments in the year after their introduction to the TF programme and considers the outcome of intervention treatment.

7 CASE STUDY PART B: TROUBLED FAMILIES ONE YEAR LATER

7.1 INTRODUCTION

This continues the analysis of the previous chapter and considers the families in the year following their introduction to the TF programme. The events that occurred for each family in the year following the start of intervention were analysed in order to determine how the family's lives had changed in that year, compared to the year before, and to determine if it was possible to understand what effect participation in the TF programme may have had upon the families. Analysis was performed on the eleven unique clusters identified in the previous chapter and on the data as a whole.

Consideration was given as to how a family's first intervention ended, whether they had further intervention treatment and the typical length of treatment. The Government guidelines as to what would constitute a family being considered 'turned around' where examined in relation to the data that was available.

Machine learning techniques, and regression methods, were performed upon the data in order to consider which (if any) factors might help to determine the likely outcome for families. That is, the techniques attempted to identify if there were any attributes that might indicate whether a family was likely to have an improvement in their circumstances, or whether their interventions might be successful.

7.2 METHODOLOGY

Analysis was performed in order to consider the complexity of events for each of the families in the year following the start of intervention, compared to the year before. A count of events occurring for each family in the year following the start of intervention treatment was compiled, to compare to the data from the previous chapter which counted events in the year prior to the start of intervention. This was performed in order to track any changes and in an attempt to detect what effect, if any, joining the TF programme had upon the families. The decision tree compiled in section 6.3.3 (Figure 49), which assigned each family to their particular cluster was utilised; the data for a family's events one year later were fed through the tree in order to determine whether a

family would stay in the same cluster or had changed cluster. Assignation to the same cluster one year later indicated that there was little change in the events that the family had; where there was a change, the new cluster assignment provided insight into the type of change that had occurred.

A count of the individual school absence records for each of the TF children before and after the start of intervention was compiled. Since the school absence data was recorded at regular intervals (each half-term), it allowed a timeline of school absence to be created for each child. Analysis and visualisations were created in order to determine whether there was any detectable change in the levels of school absence surrounding the date of the family's entry into the TF programme. Analysis was performed on the cluster-level data and on the overall group.

The Government guidelines for what constituted a family to be considered 'turned around' were analysed. There was no definitive indication within the database as to progress within the programme, or how a family's treatment was progressing. Therefore, consideration was given to how a family's first intervention treatment ended, whether they received further intervention treatment and whether the frequency of the occurrence of events changed following the start of intervention. Given this and the data that was available, an approximation of the Government guidelines was created in order to determine which of the families had some improvement following the start of intervention.

Finally, machine learning methods were utilised in order to determine which factors (or attributes) were predictors of future improvement for a family. This analysed whether particular attributes (such as, whether a family had CIN events, the cluster they were assigned to, or characteristics related to where they lived) that were known when joining the TF programme, could indicate whether a family might (or might not) have an improvement in their circumstances in the year following their entry into the TF programme. For comparison, regression models were also built, in order to determine whether they might have similar accuracy compared to the machine learning methods. The decision trees, random forests, generalized boosted models and regression models were built using the R programming environment.

7.3 RESULTS

7.3.1 Intervention Length and Further Referrals

The average length of a family's first intervention treatment was 249 days; however, 5% (115 out of 2155) of families did not have any end date for their treatment, so it was assumed that treatment was ongoing, and they were excluded from statistics pertaining to length. The length of the interventions ranged from 0 to 1503 days. Just over three quarters (77%) of families had first interventions that lasted less than a year; for 42% of families their first intervention lasted less than 6 months. It should be noted that the ECC acknowledged that the Intervention records were not all accurate: where end dates were missing, it was possible that at least some of the interventions did have end dates that were simply not logged; and particularly long interventions may not have been as long as they were logged as, some had end dates added long after the intervention had finished. However, it was not possible to determine records such as this from the available data.

Twenty-four families had first interventions that were logged as starting and ending on the same date, which would seem to imply that these families did not receive treatment. However, five of the families had treatment that was marked as a 'planned ending', implying that they did receive treatment. The remaining 19 had 'unplanned endings'. When the data was compiled all families with planned and unplanned endings were included, as families with unplanned endings had likely still received some form of treatment (even if it ended early). The families that were excluded were those who had definitely not received treatment, i.e., those who had 'inappropriate referrals' or simply had a status of 'no intervention'. Therefore, although there were a small number of families with interventions apparently lasting zero days it was felt valid to retain them in the data.

Half of the families (50%) had subsequent referrals for different types of intervention treatment, indicating that it was felt that they had more complex needs that could not be addressed by one type of intervention treatment alone. However, a referral did not mean that a family actually received any treatment; of those who had further referrals just under three quarters (74%) actually received treatment. This meant that overall, 37% of families received more than one intervention treatment; or conversely, that just under two thirds of families (63%) received no further treatment after their first intervention.

Table 24 details the percentage of families who had further referrals, and the percentage who actually received treatment, by cluster. Clusters 8 and 11, had the lowest percentage (44%) of families who were referred for further treatment, and they also had the lowest percentages of families who actually received further treatment. This may be because these families had fewer problems (just school absence, or no events at all in the year prior to intervention) than those in other clusters and so perhaps had less need for further intervention treatment. Conversely, almost two thirds (64%) of families from cluster 7 were referred for further treatment. However, overall only 40% of families from cluster 7 actually received treatment. Overall, there was a discrepancy between the percentage of families who were referred for further treatment and those who actually received it; but, as discussed previously, this was because not all referrals result in treatment (families might have been referred for treatment that was not appropriate for them, or they may simply have not wanted to participate). Cluster 5 had the highest percentage of families who actually received further treatment, with just under half (48%).

Table 24: Percentage of families who had further referrals and treatment, by cluster and overall. From ECC intervention data

Cluster	Percentage of families with further referrals	Overall Percentage of families who received further treatment
1	52%	37%
2	52%	41%
3	56%	44%
4	49%	33%
5	57%	48%
6	56%	39%
7	64%	40%
8	44%	32%
9	54%	40%
10	59%	47%
11	44%	30%
All data	50%	37%

For those families that had further treatment, just over half (54%) had further treatments that overlapped with the first intervention treatment, otherwise there was a gap between the end of the first intervention and the start of a different treatment. Table 25 details the percentage of first interventions that resulted in planned or unplanned endings, for the different intervention treatment types (this excludes incomplete, or open,

interventions). Excluding open interventions, overall, 79% of first interventions had a planned ending and 21% had an unplanned ending.

Table 25: Percentage of first interventions ending in planned and unplanned ending by treatment type, from ECC intervention data

Intervention	Planned Ending	Unplanned Ending
AO	70.1%	29.9%
CFPT	81.6%	18.4%
FF	96.3%	3.7%
FINIS	83.3%	16.7%
FIP	73.7%	26.3%

The different types of intervention clearly had differing success rates; almost all (96%) of Family First (FF) interventions resulted in a planned ending, whereas 70% of AO interventions did. Table 16, in Chapter 6, detailed the treatment types for a first intervention by cluster. Just over half of families in cluster 3 received FF treatment (as the main feature of cluster 3 was having Looked After Child events, and FF specifically targets families with this problem) and this was reflected by cluster 3 having the highest percentage of planned endings over all clusters.

7.3.2 Counting Events in the Year Following the Start of Intervention

For each family, a count was made of events occurring in the year following the first intervention start date. This was in order to track any changes and in an attempt to detect whether joining the TF programme had any effect upon the families. Whilst it could be argued that counting the events after the interventions had finished would be useful, many interventions lasted a long time, and some did not have end dates, therefore given the available timeframe of data, many families would not have been included in the 'after' analysis. Covering a continuous two-year timeframe (one year either side of the start of intervention) allowed more families to be included in the analysis, and also allowed the possibility of creating timelines of events. Much of the Government analysis also utilised this two-year timeframe. However, it should be considered (and this was also considered in the National Evaluation report (Day et al., 2016)) that one year may not be enough time to realistically tackle the problems that many of the families had.

Given the timeframes of the various data, for the 'after' analysis, only families who had a first intervention date on or before 31/07/2014 could be included, so that there was complete data for a year after this point. For families with an intervention date after this,

it would have been disingenuous to claim that they had no events simply because the data did not exist fully for a year afterwards. This left 1668 families with complete data, out of 2155, which was 74%. It should be noted that the two-year timeframes were different for all families, dependent upon when their first intervention date was.

Where a family already had a CPP or LAC event and this was still active a year after the first intervention date, this was counted (together with any new CPPs or LACs) as an event in the year following intervention. This was simply to reflect where issues were ongoing, and not imply that a family no longer had those issues. Similarly, if a family still had NEET members this was also counted.

Figure 55 plots a heatmap of the events occurring in the year following a family's referral into the TF programme (i.e. the year following the first intervention date). For comparison, the heatmap of events occurring in the year prior to the first intervention date (from Figure 19) is also included in the plot. As before, the plot provides a binary indication of each of the events (i.e. simply whether a family had that event or not), and each of the families are represented as a very thin vertical line, running from the top of the plot to the bottom (purple indicates the presence of an event, whereas turquoise indicates the absence of an event).

Whilst at first glance the two heatmaps look similar, there are significant differences between the two plots. There were proportionally more families with no events at all in the 'after' analysis; 34% compared to 28% before. The group of families with only school absence was also larger in the 'after' analysis (15% compared to 10% before). The two clusters that contained only CIN events, and only CIN events with School Absence were much smaller in the 'after' analysis, with 4% of families having just CIN events in the 'after' analysis (compared to 8% before), and 5% of families having CIN and school absence in the after analysis (compared to 11% before).

If the 'after' data had been clustered using the same criteria as was used for the 'before' intervention clustering, there would have now been only two pre-specified clusters; that is, the groups with no events, and only school absence. The groups that had only CIN events and only school absence and CIN events were no longer big enough to have been counted as a cluster.

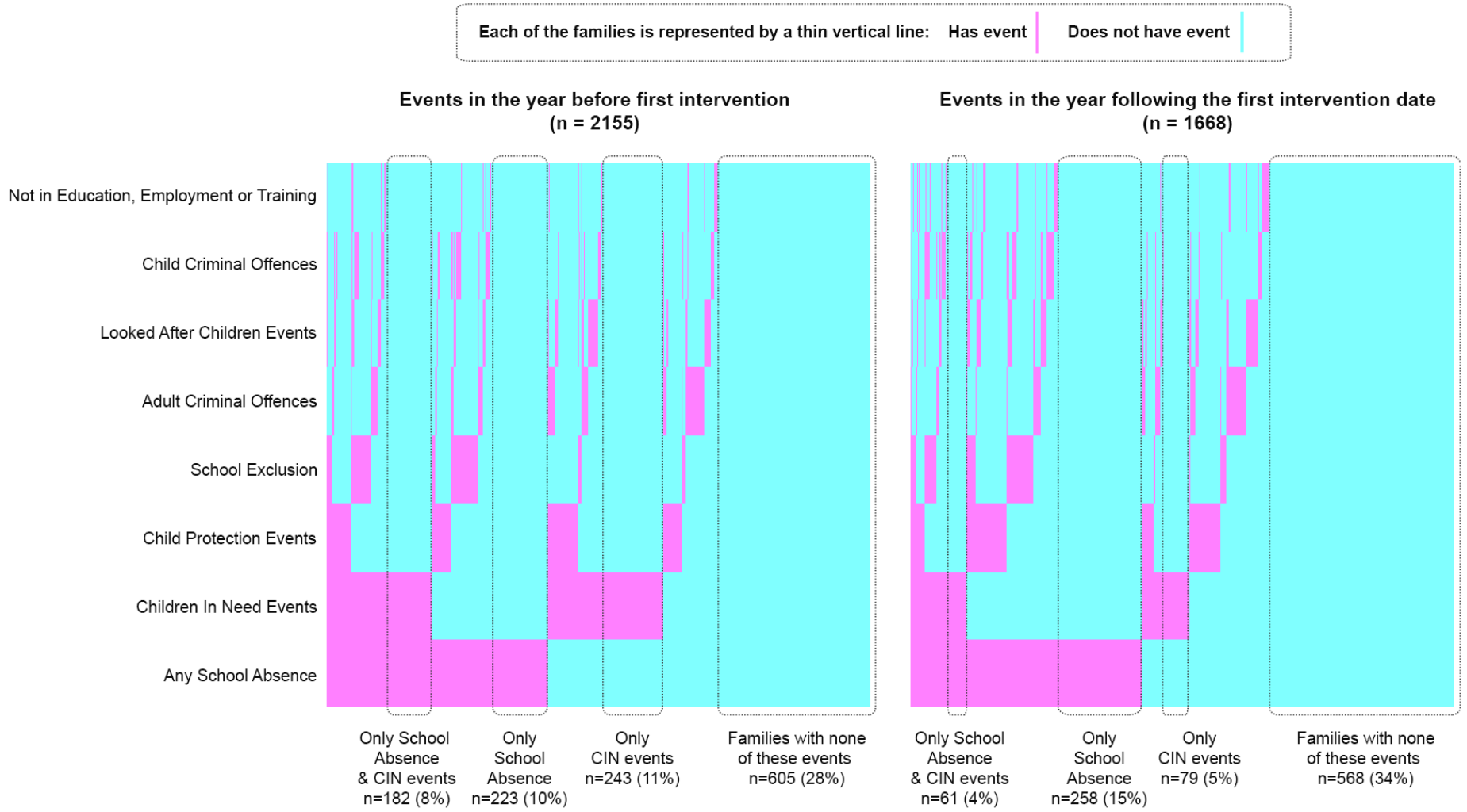


Figure 55: Comparison of events for each TF in the years prior to and following the first intervention date, utilising ECC data

Table 26 compares the percentages of families with each event in the year leading up to the first intervention date, with the percentage of families with each event in the year following the first intervention start date. Aside from CIN events (which halved), school exclusions (which stayed almost the same) and criminal offences committed by adults (which decreased slightly), the overall percentages of families having each of the individual events increased.

Table 26: Percentages of families with events in the year prior to and following first intervention date, utilising ECC data

	Year before first intervention date	Year following first intervention date
School Absence	40.6%	42.5%
School Exclusion	11.6%	11.5%
Children in Need events	40.6%	19.1%
Child Protection events	16.9%	17.5%
Looked After Children events	7.7%	9.2%
NEET	4.5%	5.8%
Adult Offences	10.3%	9.0%
Child Offences	7.5%	8.9%
DWP Benefits	42.7%	51.2%
No events	28.1%	34.7%

Overall, the decrease in the percentage of families with CIN events was most notable, and it is possible that this may be at least partially due to the way that the data was maintained and CINs were logged. Whilst CPP and LAC events had a start and end date and could therefore be considered as still continuing where applicable, the CIN events data contained very few end dates, and therefore it was treated as a one-off event (as there was no way to know if an event had ended, or even whether it was applicable for it to have an 'end'). However, it is likely that, at least in some cases, a CIN could indicate ongoing events, but the available data meant that it could not be treated this way. Therefore, for the 'after' analysis, only new CIN events could be counted, meaning that it was possible this might overlook ongoing issues. However, given that the same criterion was applied to the 'before' data (only new events were counted), the significant reduction in CIN events cannot necessarily be explained by this, but should be considered.

The national evaluation of the TF programme (Department for Communities and Local Government, 2017) found an overall reduction in children with CIN, CPP or LAC events following intervention; however, this considered the families only after they had joined the programme (at 6 months after the start of intervention compared to 12 months), and

showed relatively small changes (for CIN a reduction of 4.7%, CPP 0.4%, LAC 0.8%). Table 26 indicated that for the ECC families overall, comparing the year before to after, there were small increases in CPP and LAC events, and a large decrease in CIN events, however this is not a direct comparison, but does not necessarily match the national data. The report also noted a reduction (of 1.1%) in adults convicted of crime (in the year before compared to the year following intervention); in this case the ECC families had a similar reduction (of 1.3%).

In order to consider whether there might be underlying trends in the event data applying to the whole ECC area (as opposed to just the TF), various plots were constructed. Whilst each TF had a different timeline (their first intervention dates varied), it was felt that considering the overall trends for the whole population of the city might still provide some insight into whether any change occurred overall and not just for the TF. These plots utilised the available data, which generally covered August 2010 up to July 2015.

Figure 56 plots the monthly count of CIN events for children across the whole ECC area, with the count for just the TF children overlaid.

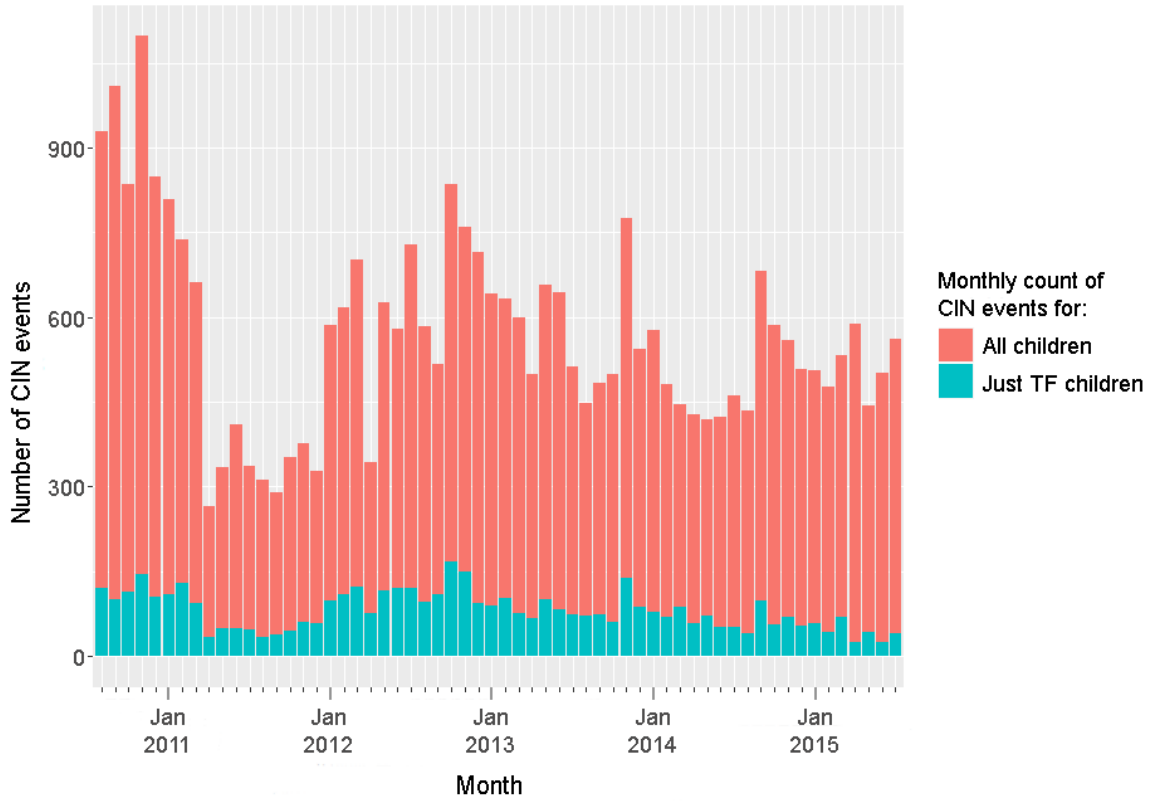


Figure 56: Monthly count of Children In Need (CIN) events for all children in the ECC area, compared to just TF children (utilising ECC data)

It shows fluctuations in the counts of CIN events over time, and also that the TF CIN count tends to follow the trend of the data for the overall population. However, most significantly it illustrates a large drop in CIN events from April 2011 up to December 2011. It was considered that this might help explain the difference in prevalence of CIN events 'before' and 'after' intervention treatment for the TF, however it did not. Most families did not start their 'after' time period in this timeframe (and so could not be affected by it), and the few (6%) of families that did, had CIN events both 'before' and 'after' and so were not in the group who had no further CIN events. TF children accounted for, on average, 14% of all CIN events.

Similar plots were created for Child Protection (CPP) events (Figure 57) and Looked After Children (LAC) events (Figure 58). In particular, the plot for CPP events highlighted that the TF children accounted for a large proportion of the overall CPPs in the ECC area, on average 32%. It seems also that the general trend for the TF was of increasing numbers of CPPs (albeit with a dip in the second half of 2013), followed by a gradual decrease towards the end of 2014.

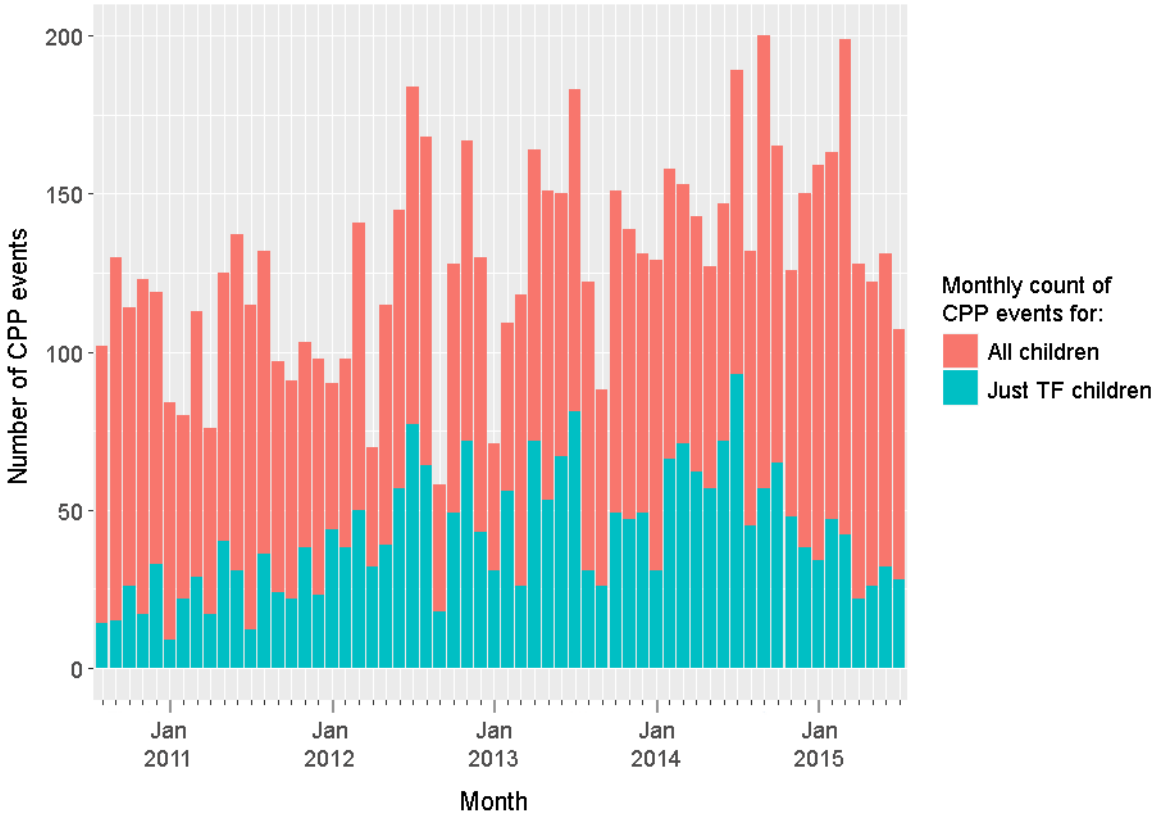


Figure 57: Monthly count of Child Protection Plan (CPP) events for all children in the ECC area, compared to just TF children (utilising ECC data)

The LAC data (Figure 58) also appeared to follow a similar trend to the CPP data for the TF, with a pattern of gradually increasing events, followed by a drop towards the end of 2014. However, in contrast to the CPP events, TF children accounted for a smaller proportion of the overall LAC events (on average 19%).

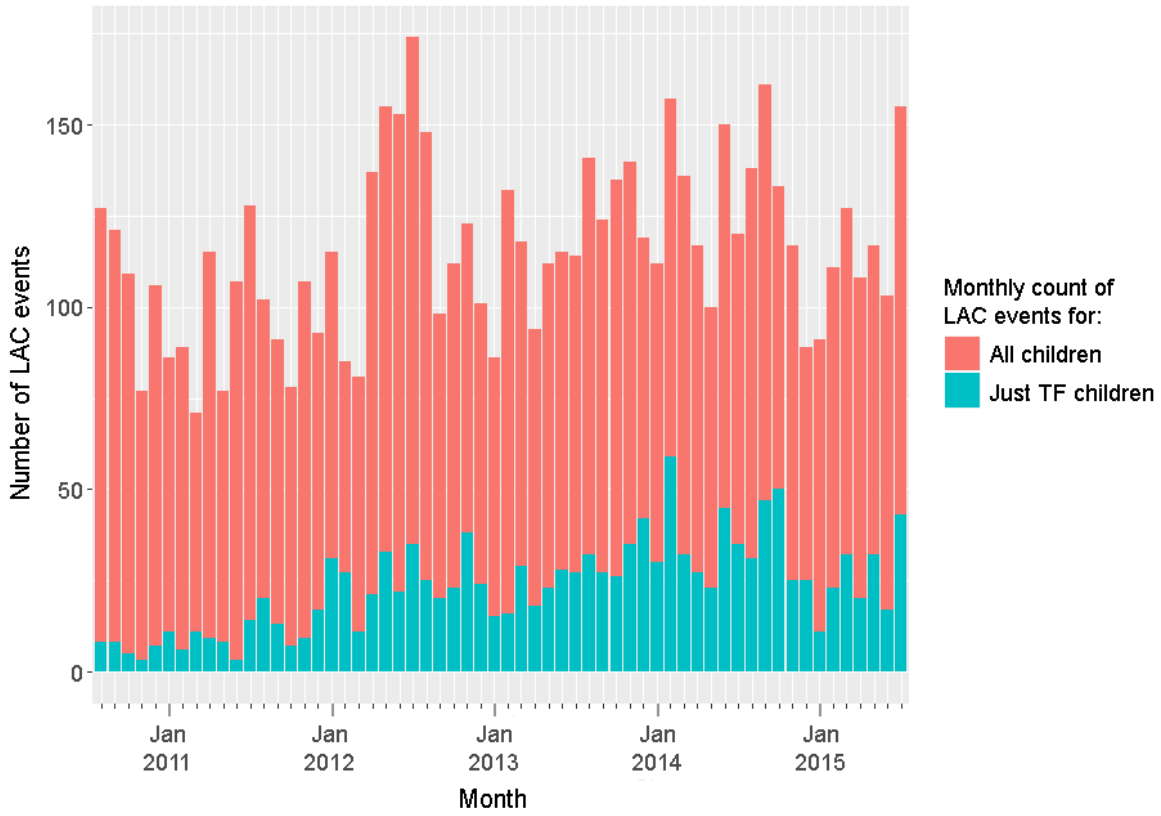


Figure 58: Monthly count of Looked After Children (LAC) events for all children in the ECC area, compared to just TF children (utilising ECC data)

Figure 59 plots the average percentage of unauthorised school absence for all pupils in the ECC area, compared to just the TF pupils. This was calculated for each half-term for the school years 2010/11 to 2014/15. The overall trend, for all pupils, stayed fairly constant, albeit with fluctuations; the mean unauthorised absence for the whole time-period was 1.6%. However, the general trend for just the TF pupils was of increasing school absence, though again with fluctuations.

Figure 60 plots the count of school exclusions, for each half-term, for all pupils in the ECC area compared to just the TF pupils. The general trend of the TF pupils follows the fluctuations of the overall data, but shows more of a constant level (than the decreasing trend overall in the first two school years). TF pupils accounted for, on average, 13% of all school exclusions.

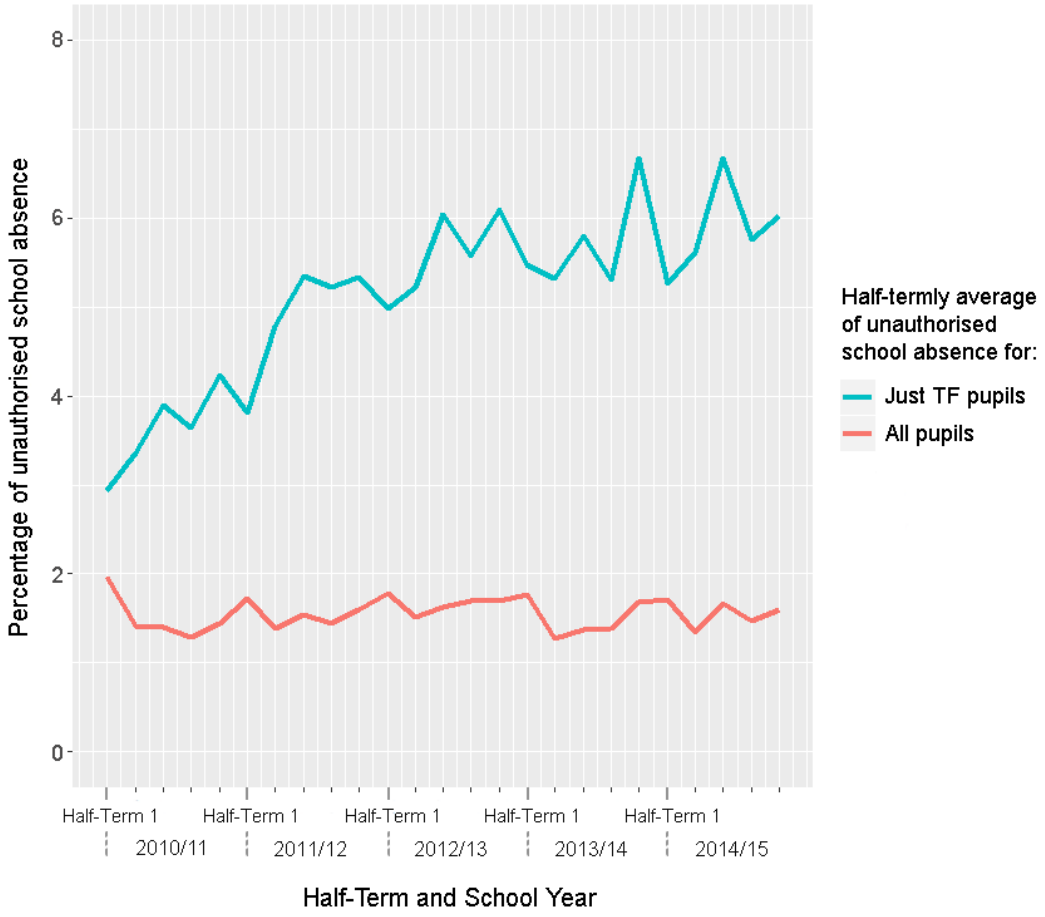


Figure 59: Half-termly average percentage of unauthorised school absence for all pupils in the ECC area, compared to just the TF pupils (utilising ECC data)

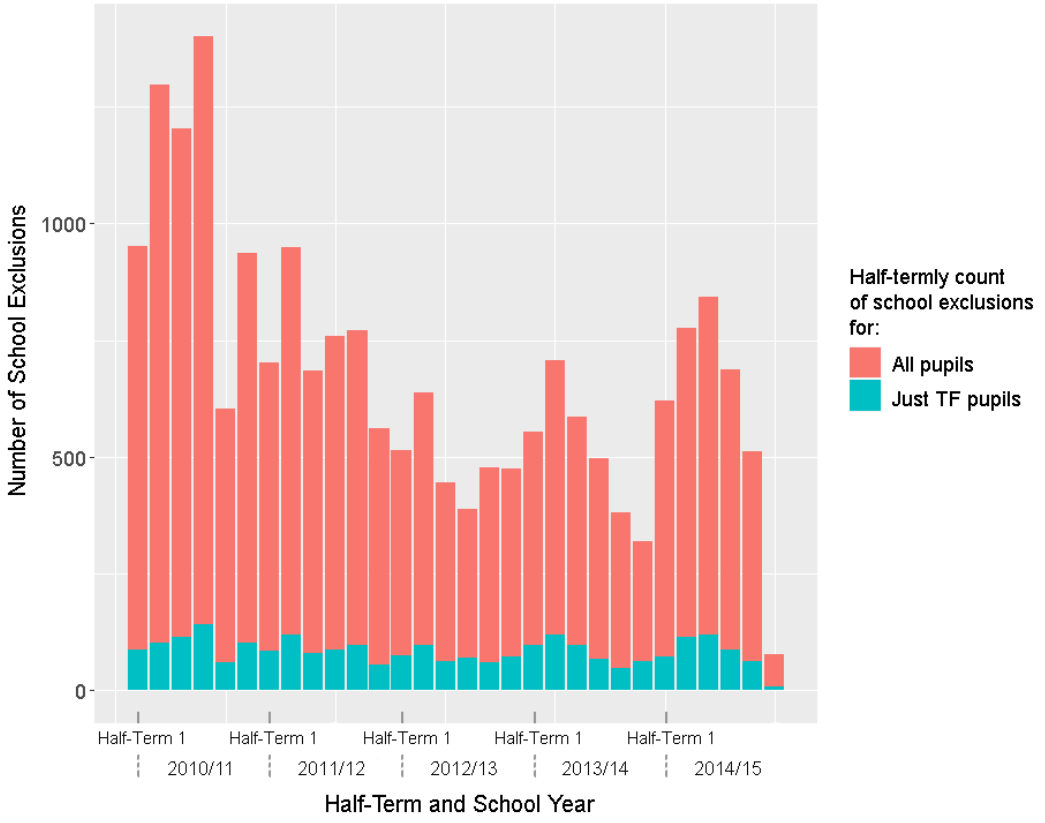


Figure 60: Half-termly count of school exclusions for all pupils in the ECC area compared to just the TF pupils (utilising ECC data)

Figure 61 plots the monthly count of individuals logged as Not in Employment, Education or Training (NEET); this counts only new incidences and not ongoing ones. It shows a gradually increasing trend overall, with large fluctuations. The TF generally follow this trend and account for, on average, 7% of all new NEET incidences.

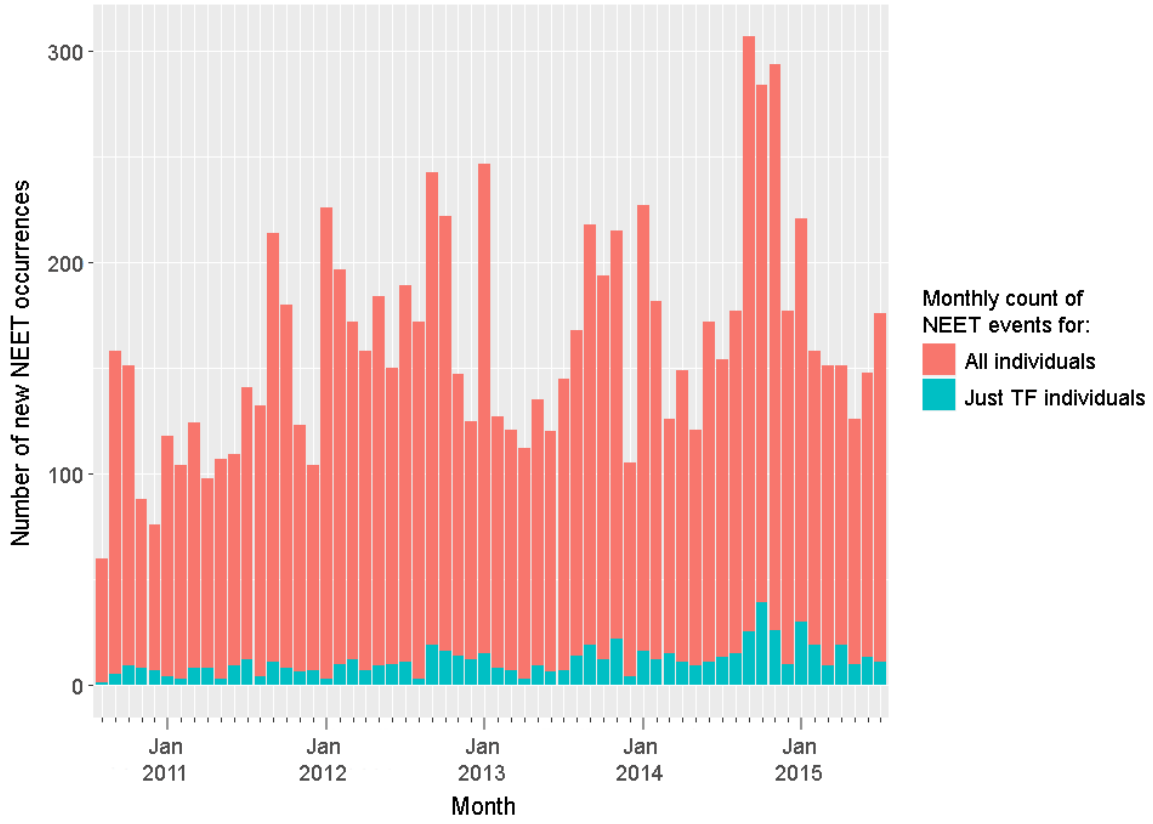


Figure 61: Count of NEET incidence per month, for all individuals in the ECC area compared to just TF individuals (utilising ECC data)

Figure 62 plots the monthly count of criminal offences committed by adults for the ECC area, together with those committed by TF adults. In contrast to the other plots, the TF account for a much smaller proportion of the crimes committed overall, on average 3%. There is a clear trend overall, of the counts of crime reducing month by month. The counts of crime for TF generally stayed more constant and exhibited less of a reduction.

Figure 63 plots the monthly count of criminal offences committed by children (aged under 18) for the whole ECC area, together with those committed by just TF children. In contrast to the plot detailing crimes committed by adults, it shows that TF children were responsible for a higher proportion of crimes, on average 11%. However, similar to the adult crimes plot, it highlights that the overall trend was of decreasing numbers of crimes being committed by children. The crimes committed by TF children did not follow this

decreasing trend, and instead increased slowly initially, stayed somewhat level, then dropped a little towards 2015.

Figure 62: Monthly count of criminal offences committed by adults for the ECC area, compared to just TF adults (utilising ECC data)

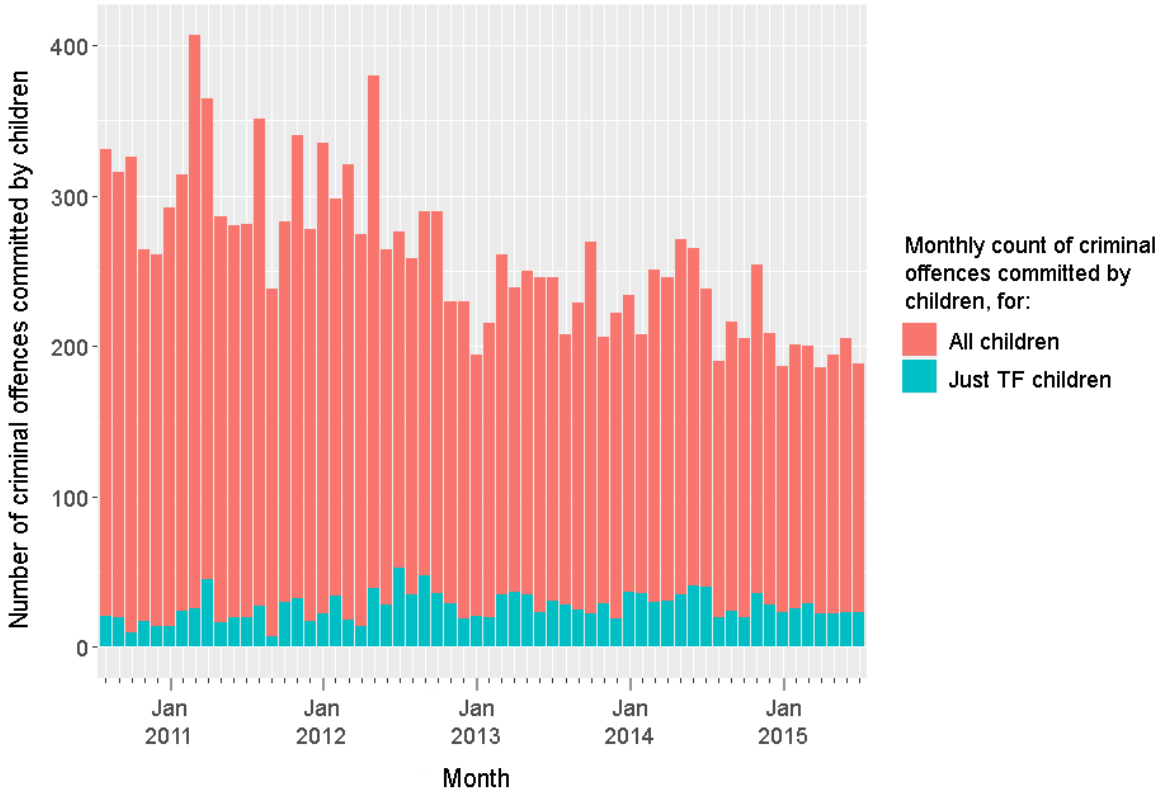


Figure 63: Monthly count of criminal offences committed by children (aged under 18) for the ECC area, compared to just TF children (utilising ECC data)

Whilst the plots of trends are useful in providing the context of the underlying data, it should be considered that each family had a different timeline (that is, their first intervention dates were all different), and so each of their two-year timeframes would be located at different points on these plots. However, the plots provide overall insight. The loose overall trend for CPP, LAC and NEET events was that of fluctuating increase for the whole population; referring back to Table 26, these events for the TF also had small increases and so might be on trend. Similarly, the overall trend for crimes committed by adults was decreasing, and the TF data also reflected a small decrease. However, the overall trend for crimes committed by children was decreasing, yet the TF data reflected a small increase.

Table 27 details the percentage of families with each event in the years before and after the first intervention date for each cluster.

Table 27: Percentage of families in each cluster with events in the year prior to and following first intervention date (with interesting percentage highlighted in bold). Utilising ECC data

Cluster	Absence		Exclusion		CIN		CPP		LAC		NEET		Adult Offences		Child Offences	
	Before	After	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.
1	66	63	57	32	42	29	1	15	0.3	9	0	7	36	14	27	23
2	44	46	6	10	58	14	**100**	**38**	12	18	2	4	17	16	4	4
3	30	42	11	14	63	16	2	32	**100**	**42**	1	1	21	9	8	8
4	43	47	16	16	31	29	0	16	10	10	**100**	**63**	15	16	21	18
5	29	36	24	21	10	21	5	0	0	0	0	0	**100**	**36**	10	7
6	100	88	54	30	57	22	22	25	9	8	15	25	13	8	43	30
7	44	70	24	39	32	30	44	17	4	22	0	13	8	4	100	83
8	100	79	0	7	0	22	0	8	0	4	0	2	0	4	0	9
9	0	19	0	7	**100**	**28**	0	16	0	4	0	1	0	4	0	5
10	100	87	0	12	**100**	**22**	0	28	0	7	0	6	0	6	0	8
11	0	9	0	1	0	11	0	6	0	2	0	3	0	6	0	0.4
All data	41	43	12	12	**41**	**19**	17	18	8	9	5	6	10	9	8	9

For those clusters where all families had a particular event, that is, the value was 100%, specifically CIN (clusters 9 and 10), CPP (cluster 2), LAC (cluster 3) and NEET (cluster 4), there was a significant decrease in the percentages of families with those events in the year following the start of intervention. The drop in the percentage of families having CIN events was most notable (clusters 9 and 10). However, for school absence (clusters 6, 8 and 10) and child criminal offences (cluster 7) there was a much smaller decrease. For

many of the clusters, where in the before analysis there was a zero for a particular event (for example, families in cluster 8 had no school exclusion before the start of intervention), in the 'after' analysis there was an increase (e.g. 7% of families in cluster 8 had exclusions in the year after the start of intervention).

Tables 26 and 27 highlight one of the main benefits of identifying the different clusters in the data; if just a 'global' overall analysis had been performed, then it showed that in general there was a small increase (aside from CIN events) in the percentage of families having most of the events in the year after the start of intervention compared to before. However, considering the clusters separately provided context; the results showed that where considered on the cluster-level, in many cases there was a decrease in the percentage of families having the most important events for that cluster.

Table 28 considers attributes that were contained in the data but were not clustered upon; for each cluster, it compares the percentages of families who had the particular events in the year prior to and following the start of intervention. As previously stated, the DWP benefits data had missing data, and the address data may also have been unreliable, nevertheless they were included for comparison, but with these caveats. The Drug/Alcohol and Domestic Abuse data were thought to be reliable but were not clustered upon because they were a subset of the CIN, CPP, and criminal offences data; however, they could still provide useful contextual information.

Table 28: Percentage of families in each cluster with events not clustered upon in the year prior to and following start of intervention (with interesting percentages highlighted in bold). ECC data

Cluster	Receiving DWP benefits		Changed address at least once		Drug/Alcohol Events		Domestic Abuse Events	
	Before	After	Before	After	Before	After	Before	After
1	48%	55%	46%	33%	2%	4%	14%	11%
2	46%	50%	53%	38%	5%	4%	17%	7%
3	35%	45%	**73%**	**34%**	2%	1%	7%	1%
4	**57%**	**78%**	49%	33%	5%	6%	5%	8%
5	57%	71%	48%	29%	**0**	**14%**	**33%**	**14%**
6	57%	65%	**48%**	**55%**	4%	5%	6%	8%
7	28%	35%	64%	48%	0	0	**20%**	**9%**
8	42%	46%	30%	23%	1%	1%	0	1%
9	36%	49%	54%	28%	3%	2%	13%	6%
10	42%	51%	42%	33%	4%	6%	14%	6%
11	40%	51%	36%	25%	1%	2.%	0	4%
All data	43%	51%	45%	31%	3%	3%	8%	6%

After the start of intervention, there was an increase in the percentage of families receiving DWP benefits for all clusters, with cluster 4 having the largest increase.

However, since as previously, discussed there were concerns about the accuracy of the DWP benefits data (not all historical records were retained) it is difficult to draw firm conclusions from this. It may be that the missing data distorts these statistics (for instance, that there was a greater amount of older data missing than newer and hence the 'before' analysis was missing more data than the 'after' analysis), or it may be that entry into the TF programme corresponded with more families applying for state benefits (perhaps, once families had a dedicated key worker, they were given more assistance in applying for the various benefits that they might qualify for).

In contrast, aside from cluster 6, the percentage of families who had changed address at least once decreased in the year after intervention. Cluster 3, in particular, had a large decrease in the percentage of families who had changed address after the start of intervention, compared to before. Cluster 3 consisted of families who all had Looked After Children events before the start of intervention. However, in the year following intervention, only 42% of families had LAC events; it is possible that the reduction in families with LAC events also corresponded with the decrease in changes of address (as if fewer families had children in care, there would be fewer address changes logged). Over all the clusters, the decrease in address changes may cautiously indicate greater stability in the lives of some of the families following the start of intervention treatment.

The percentage of families with Drug/Alcohol events had very little change in the year following the start of intervention, although families in cluster 4 went from having none before intervention to 14% of families after the start of intervention. The percentage of families with domestic abuse events decreased for all but four of the clusters, in particular cluster 5 had the largest decrease (of 19%), and clusters 7 and 2 also had decreases of 11% and 10%.

In order to compare the cluster characteristics from the year before intervention to the cluster characteristics of the families a year later, Figure 64 contains two Nightingale plots: the plot on the left contains the original cluster characteristics derived from events in the year before intervention; the plot on the right contains the cluster characteristics for the families in the year following the start of intervention. It must be noted that whilst the Nightingale plot provides a clear visual aid of the important attributes in each cluster, it is very much a tool for comparing cluster characteristics within the particular group of 11 clusters. Direct comparison of, for instance, cluster 1 'before' and cluster 1

'after' does not necessarily make sense. This is because the size of the coloured segments is derived by considering all 11 clusters together. For instance, cluster 7 on the 'after' plot has a full yellow segment, indicating that of all the clusters, cluster 7 had the highest proportion of families with CIN events. The 'before' plot for cluster 7 had a much smaller yellow segment because six other clusters had higher proportions of families with CIN events. However, when the actual statistics are considered, families in cluster 7 had a decrease in CIN events overall (from 32% 'before' to 30% 'after'), yet when comparing the two, one would assume that there was a large increase. This highlights that a direct comparison can be dangerous. However, the two plots are considered together in order to highlight the changes that occurred overall and to provide a visualisation of the overall change in patterns after one year.

The plots indicate that the families in cluster 11 went from having no events to having a small level of events after the start of intervention. In many cases the clusters retained the primary characteristics (cluster 4 still had the highest percentage of families with NEET members, cluster 6 still had the highest absence levels over all the other clusters, and cluster 5 still had the highest proportion of families with adults who had committed criminal offences. However, as Table 27 highlighted, these clusters all had great decreases in the percentage of families having those events after the start of intervention.

Figure 64: Nightingale plot comparison of cluster characteristics in the year before and after the start of intervention (Using ECC data)

7.3.3 Comparison of cluster assignments one year later

Table 27 and Table 28 highlighted that there were changes in the types of events that families in each cluster had in the year following the start of intervention compared to before. To determine the scale of these changes in terms of the cluster assignments, the data compiled for the year following the start of intervention was fed into the decision tree model (Figure 49) developed in the previous chapter. This utilised the decision tree rules to assign each family to one of the 11 cluster types, using the data compiled for the year following the start of intervention. This analysis was performed in order to determine whether a family still belonged in their original cluster a year later (i.e. that the type or frequency of their events had not changed significantly) or whether they now belonged to a different cluster (i.e. that there had been a sufficient change in the type or frequency of events in the year following the start of intervention to mean they no longer belonged in their original cluster).

An Alluvial plot was created, Figure 65, to indicate which clusters the families were assigned to at the start of intervention, and which they were assigned to one year later; it highlights the changes from cluster to cluster. On the left of the plot are the original cluster assignments that each family received, utilising the data compiled in the year before their first intervention date. On the right side of the plot are the cluster assignments for each of those families utilising the data compiled in the year following the start of their first intervention date. This analysis includes only the 1668 families that had complete data for the year following the start of intervention. On either side of the plot are the percentages of families in each cluster.

Most notable in Figure 65 is that almost three quarters (73%) of the families assigned to cluster 11, were still in cluster 11 a year later. Cluster 11 contained families that had none of the specified events, therefore this meant that these families had no events in the years prior to and after the start of intervention treatment. Cluster 11 grew in size after the start of intervention (it initially contained 28% of families and this increased to 34% a year later), and it received families from each of the other ten clusters. Almost half (45%) of cluster 9 (which contained families with just CIN events) were assigned to cluster 11 a year later, meaning that they went from having just CIN events to having no events. And just over a third (36%) of cluster 5, and a fifth (20%) of cluster 2 moved to cluster 11.

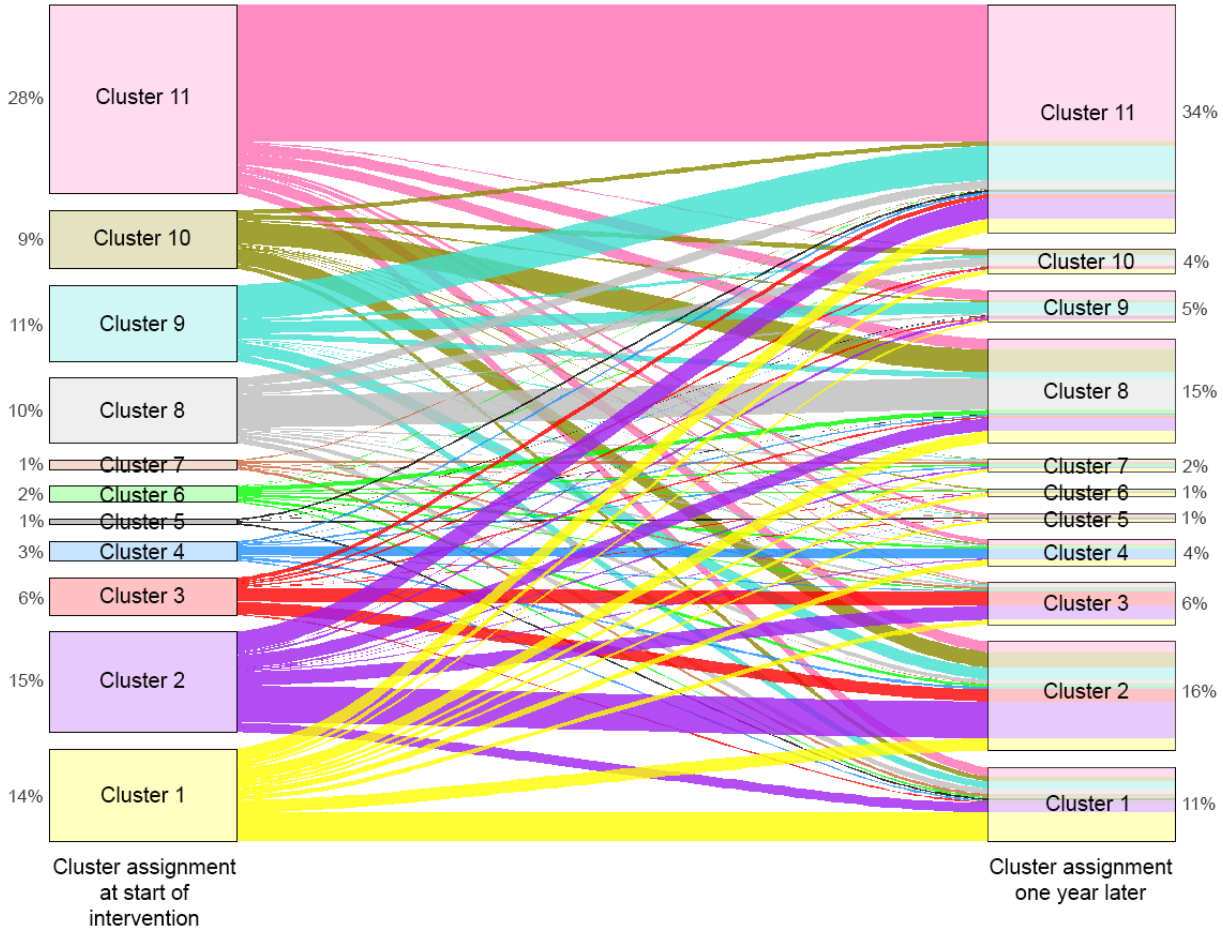


Figure 65: Alluvial plot showing the change in cluster assignments one year after the start of intervention

Table 29 details the percentage of families from each cluster who had no further events after the start of intervention (i.e. they were assigned to cluster 11).

Table 29: For each cluster, the percentage of families who had no further events after the start of intervention treatment

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Percentage of families who had no further events after the start of intervention	15%	20%	12%	10%	36%	3%	9%	13%	45%	8%	73%

Considering Figure 65, aside from cluster 11, clusters 8 (only families with school absence) and 4 (families with NEET members) had the highest percentage of families who stayed in the same cluster a year later (that is, nothing changed significantly), with both having just under half of families remaining (49% and 47% respectively). In contrast, cluster 10 (which contained families with just school absence and CIN events) had only 8% of families remaining a year later. The majority of cluster 10 were assigned to cluster 8 (39%, just school absence) and cluster 2 (27%, families with Child Protection Plans). This would appear to mean that the families from cluster 10 who moved to cluster 8 had an

improvement in circumstances, whereas those who moved to cluster 2 had a decline in their circumstances, as moving from having just CIN events to having a CPP (Child Protection Plan) implies an escalation in child safeguarding issues.

Just over a third (36%) of families from cluster 3 (those with Looked after Children) remained in cluster 3 a year later; however, another third (32%) changed to cluster 2 (meaning they had a Child Protection Plan issued in the year following intervention). This may imply that there was an improvement in these family's needs since moving from having a child in care to having a CPP would seem to be an improvement.

Whilst the alluvial plot may look a little complex, as plotting eleven clusters (and eleven different colours) produces a somewhat messy image, it does give an overview of the type of changes that occurred in the year following the start of intervention. It highlights the increase in families with no events (cluster 11), and those with just school absence (cluster 8); and the decrease in families with just CIN events and just school absence and CIN (clusters 9 and 10).

7.3.4 School Attendance Timelines

Whilst for most of the events, the analysis consisted of simply counting how many occurrences there were (or whether there was any occurrence) before and after the start of intervention, for the school absence data it was possible to create a timeline. School attendance data was collected every half term, which meant that for each child the data could be compiled to create a picture of the trend of their school attendance. The half-termly data indicated how many school sessions a child attended, and how many were actually available to them; from this, the percentage of unauthorised absence was calculated.

The half term during which a child's first intervention began was treated as half-term 0; if their first intervention date was between half-terms (for example, in the summer holidays), then the following half-term was chosen as half-term 0. Absence data for half-term 0 and the five half-terms before and after this point were compiled, which covered the 'before' and 'after' the start of intervention timeframe. In the ECC data, a school year was represented by five half-terms, as data was only available for the first five half-terms of each school year (there would normally be six). This was standard across England; absence data for half-term 6 only began to be collected from the 2013/14 school year

onwards. However, the ECC data did not receive this data until after the data for this case study was collected.

Approximately 70% (2772 out of 3970) of the TF children were of school age (aged between 5 and 16) a year before the first intervention start date, which was where the timelines began. However, a run of eleven consecutive half-terms worth of attendance data was required for the analysis, and this filtered out a portion of children. Complete data was compiled for 1092 children, which was approximately 39% of the school-aged children overall. Many children had some school attendance data, but it did not cover a consecutive run of half-terms. It was not clear why there were so many gaps in the school absence data; reasons could be children who had moved in and out of area or changed schools, or simply missing data. However, children from each of the eleven clusters were represented in the data, as detailed in Table 30, although there were particularly small numbers of children from clusters 3 to 7 represented.

Table 30: Number and percentage of children with absence timeline data from each cluster, utilising ECC data

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Number of children	230	200	32	40	5	33	17	201	56	181	86
As a percentage of school age children in cluster	47%	44%	21%	32%	24%	32%	35%	49%	21%	52%	24%

Figure 66 visualises the timelines of all 1092 children. The plot shows a chaotic picture, which made it difficult to identify any patterns. However, when the attendance timelines were plotted by cluster (Figure 67) and by the overall trend for each cluster (the average percentage of unauthorised school absence, in Figure 68), it was possible to identify patterns.

Overall, the chaos in Figure 66, where no discernible pattern is present, compared to the more understandable plots in Figure 67, highlight the advantage of analysing the data on the cluster level as opposed to one overall global analysis.

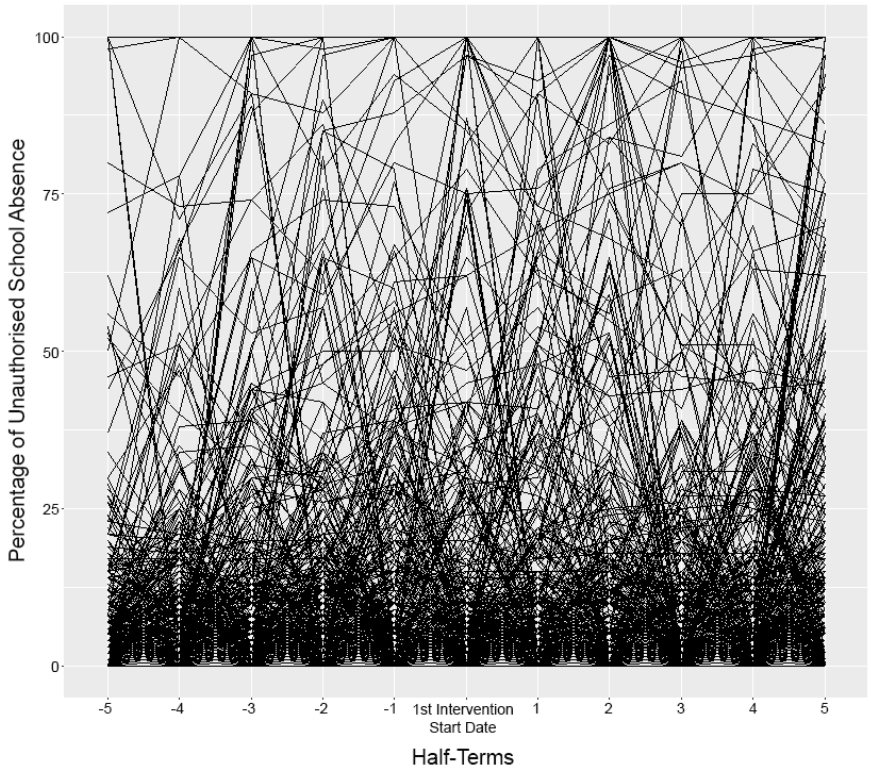


Figure 66: Timelines of school absence for the five half-terms before and after the start of intervention for all applicable children, ECC data

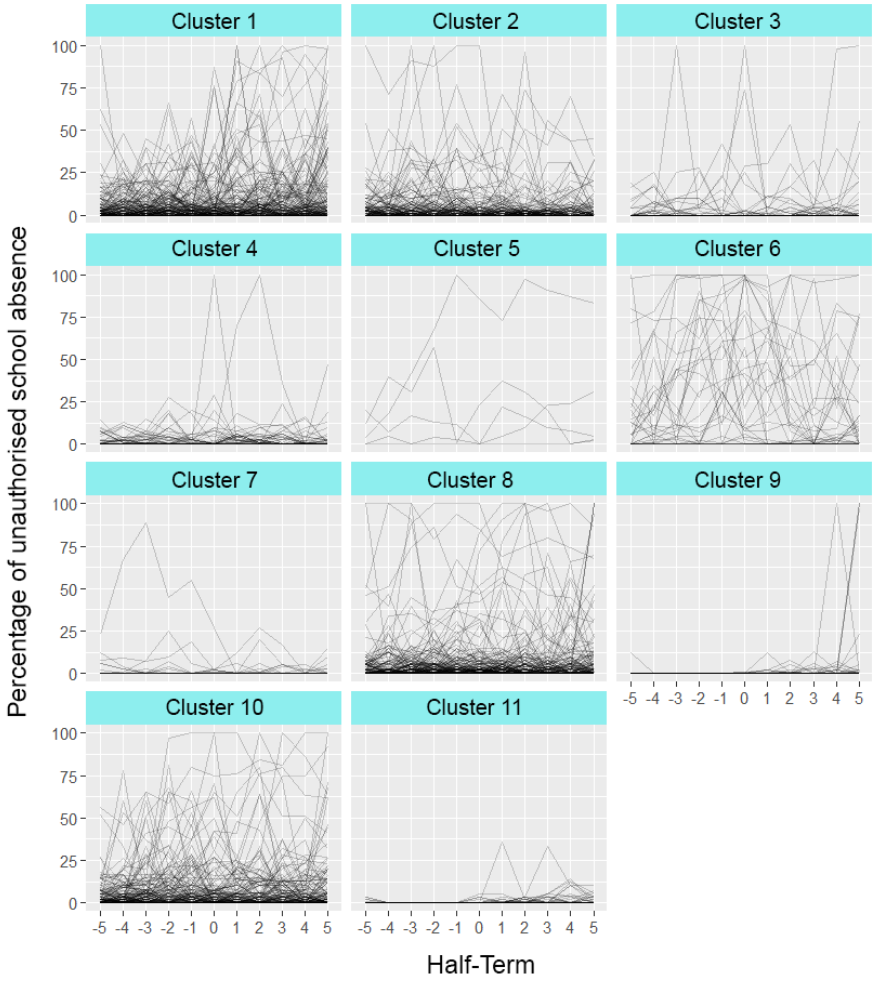


Figure 67: Individual school absence timelines for the five half-terms before and after the start of intervention for children in each cluster, ECC data

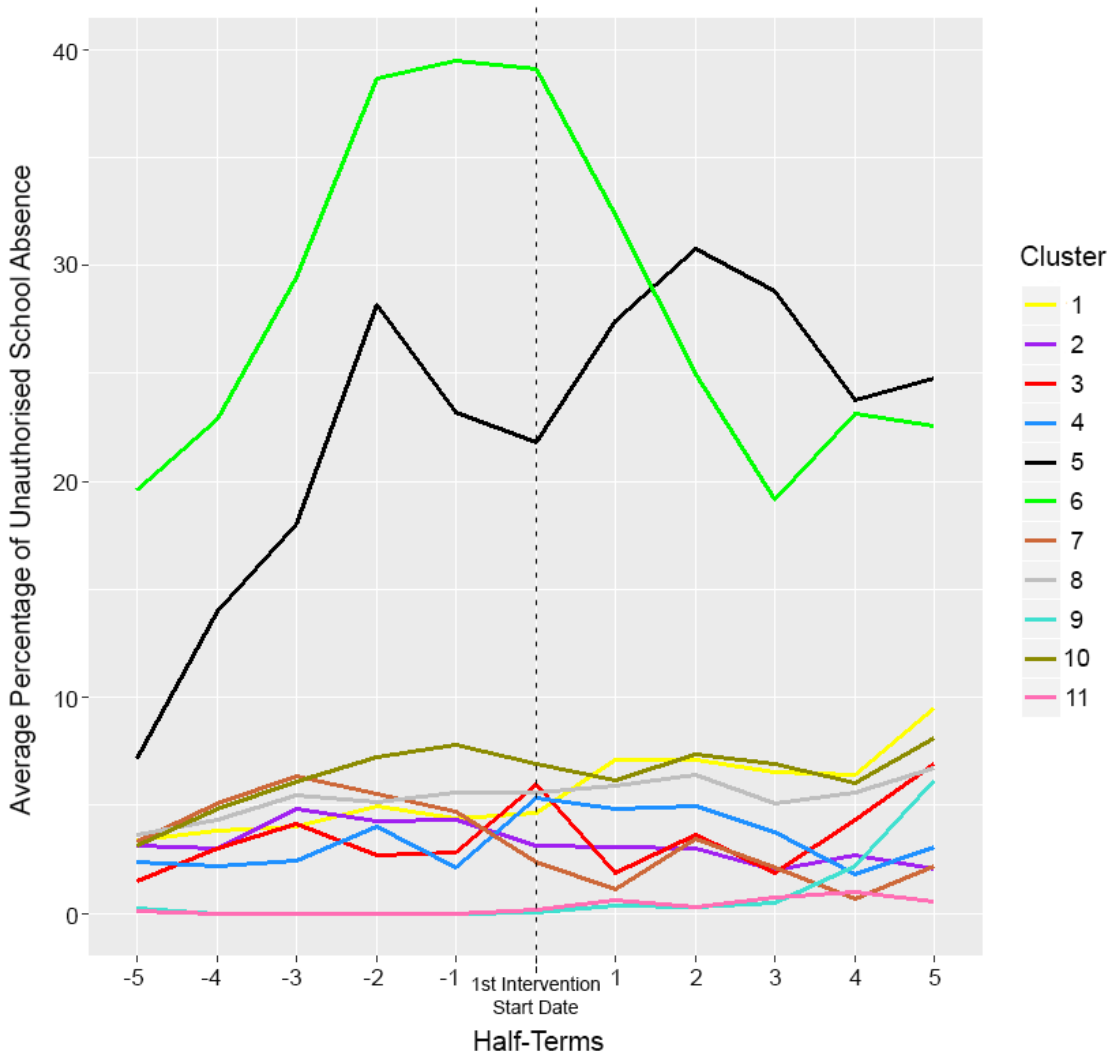


Figure 68: Average percentage of School Absence by Cluster, for the five-half terms before and after the start of intervention, ECC data

Where considering the average percentage of school absence by cluster (Figure 68), the plot shows that for cluster 1, absence levels increased overall after the start of intervention, and this is reflected by the slight upwards slope. For cluster 2, the average absence levels levelled out where intervention began, but overall the trend was decreasing average absence, starting 3 half-terms before the beginning of intervention.

Clusters 9 and 11, which had no school absence before intervention (represented by flat lines), had a small increase after the start of intervention. Whereas the overall trend of cluster 11 was a very slight increase, for cluster 9 absence increased sharply from term 3 onwards. However, cluster 9 was represented by a relatively small sample of children (56).

Clusters 8 and 10, which contained families who all had some school absence, both had increasing average absence before intervention. For cluster 10 this began to decrease

one half-term before intervention and then began to fluctuate one half-term after intervention. Cluster 8's average absence decreased a little two half-terms after intervention, but then rose again.

It should be noted that clusters 3 to 7 were each represented by a small number of children, and so any trends may not be as reliable as for the larger samples. However, cluster 3 had an overall fluctuating trend, with an increase one half term before intervention, and then a decrease at the start of intervention. Cluster 4 had an overall increasing trend before intervention, which then decreased following the start of intervention.

Clusters 5 and 6 were most notable, with the highest levels of average school absence. They also both had what appeared to be a sudden change in absence levels corresponding to the start of intervention (a sharp drop for cluster 6 and a sharp rise for cluster 5). However, they both also had a noticeable change in absence levels two half terms before intervention, and then again two half terms after the start of intervention. Cluster 5 was represented by a very small sample of children (n=5) and had a fluctuating pattern, but overall, average absence levels increased over the two years. Cluster 6, which had a slightly larger sample of children (n=33) had a noticeable drop in absence levels at the start of intervention, lasting until half term 3. Of all the clusters, cluster 6 is the one that had the most change corresponding with the start of intervention treatment.

Overall, whilst there were differences in the particular timelines for each cluster, it was difficult to attribute them to the start of intervention. There were certainly changes around the start of intervention for some clusters, but this was also true after and before. It was also perhaps unlikely that any changes would occur immediately, it may be that if changes or improvements were likely to occur that these would follow at some point after the start of intervention (as it was perhaps unlikely that intervention treatment would have an immediate effect). The overall Government report (Department for Communities and Local Government, 2017) into the programme in England also noted that school absence generally fluctuated overall, and that it rose and fell in the terms following the start of intervention. It would seem therefore that the ECC data followed a similar trend to the national data.

As an alternative consideration, Figure 69 plots the overall trend of the school absence timelines when grouped by the OFSTED rating for each child's school. The OFSTED rating

to some degree could be thought of as a geographical, or 'place-based' attribute, since, in general, children attend schools that they live close to. The sample sizes of each group were: 113 children attended 'Outstanding' schools; 576 attended 'Good' schools; 231 attended schools that 'Required Improvement'; and 33 attended 'Inadequate' schools. It was noted that the group for children attending 'Inadequate' schools was small; however, this was likely to be the case as, in general, fewer schools are classed as inadequate.

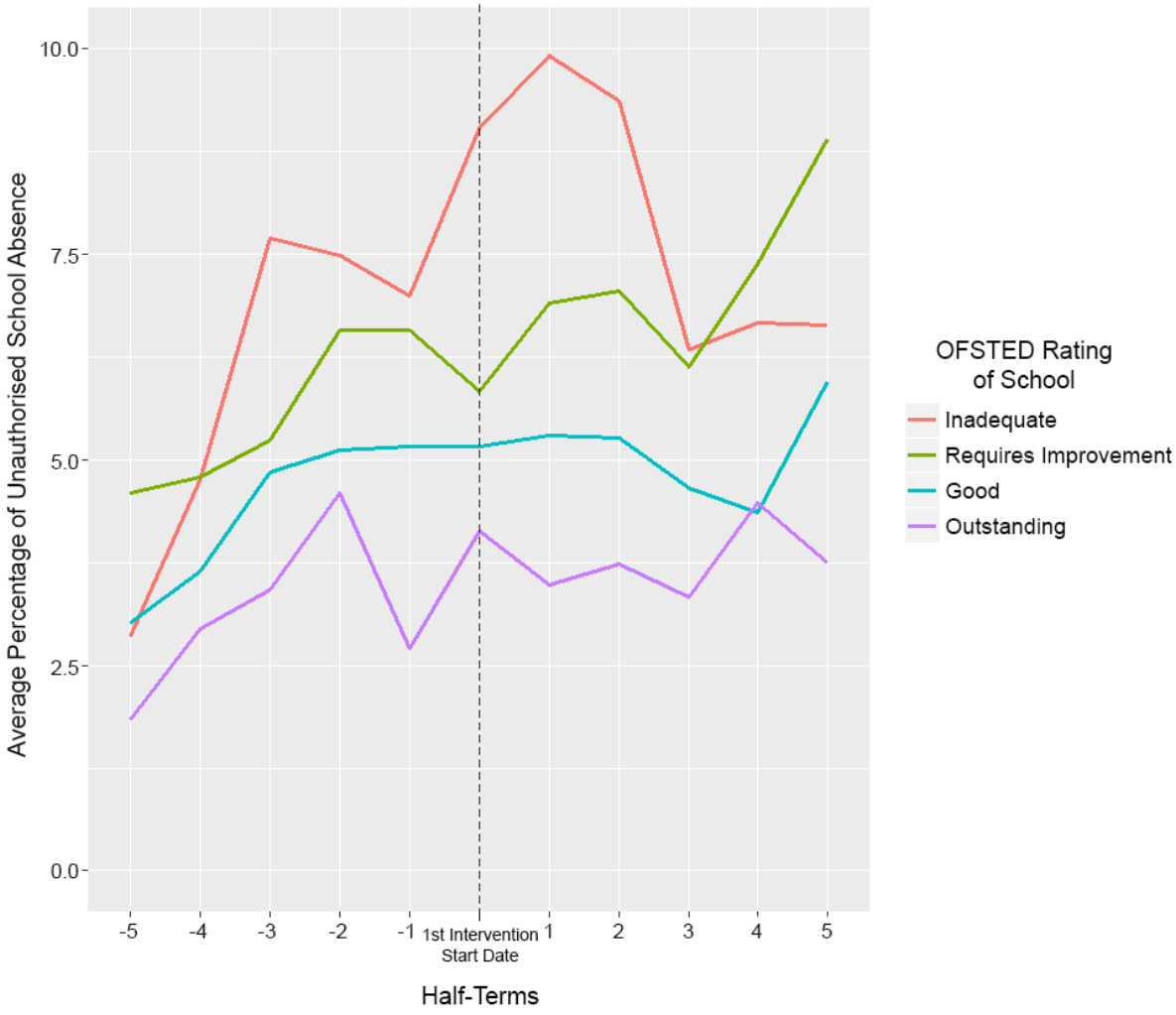


Figure 69: Absence timelines aggregated by school OFSTED rating, for the five half-terms before and after the start of intervention (ECC data linked to Department for Education (2016) data)

The plot, Figure 69, indicates that children attending schools ranked as 'Outstanding' had the lowest average levels of school absence in the timeframe surrounding their first intervention. And that the worse the OFSTED rating of the child's school was, the higher the levels of average school absence were. It was difficult to determine what effect, if any, intervention might have had; however, for those attending 'Inadequate' schools there was a noticeable drop in absence levels one half term after the start of intervention.

In general, all four groups had increasing average absence levels prior to the start of intervention, however they all decreased (or slowed down) between two and three half terms before the start of intervention, therefore this would not appear to be an effect that could be attributed to the TF programme.

As with the overall absence data, from the plots alone it was difficult to determine if starting intervention treatment had an effect upon the school absence for individuals in the families. The plots did highlight the overall trend for each cluster, and in particular the higher levels of absence for clusters 5 and 6. Grouping the families by school OFSTED rating was interesting as it indicated that, at least for the TF, average school absence levels followed the school rating; the better the rating the lower the unauthorised absence.

7.3.5 Considering the Families One Year Later

Analysis was performed in order to determine what change (if any) had occurred in the year following a family's introduction to the TF programme. There was no indication in the database where a family was considered to have been 'turned around', or even where a family was no longer in the TF programme or likely to still be receiving help. There was no indication even of when a family might have left the area. The only way to determine progress (or whether a family was still receiving treatment) was to analyse the events that occurred for them, and to consider whether they were still receiving intervention treatment.

This analysis includes only the 1668 families for whom a years' worth of data existed after the start of intervention, and it considered the Government guidelines of what constituted a family being 'turned around'. For both Phase 1 and 2, having an adult in the family who had moved off out of work benefits and into employment meant that the family could be considered 'turned around'. However, as previously discussed, this particular information on state benefits was missing from the database, therefore evaluation of this criterion was not possible.

The other criteria for Phase 1 specified that a family could also be considered 'turned around' where each child had fewer than 3 school exclusions and less than 15% school absence, there was a 60% reduction in anti-social behaviour for the whole family, and the offending rate for all children was reduced by at least a third. Analysis was performed using these criteria, however since there was no anti-social behaviour data available, only

school exclusion, absence and criminal offences committed by children could be considered. And since these events pertained only to children, it meant that families without children were not considered (since having no school absence, etc. was meaningless for families that consisted of only adults). Table 31 details the number and percentage of families from each cluster that met this reduced Phase 1 criteria for being 'turned around' (that, is all children in the family had: school absence less than 15%; fewer than 3 school exclusions; and a reduction in criminal offences (or none at all) in the year following the start of intervention). It should be noted that this simply considered whether the count of events for each child fell under this threshold in the 'after' data (they may have had none of these events 'before').

Table 31: Phase 1 reduced criteria - number of families whose children met the criteria in the year following the start of intervention, by cluster (with percentages in parentheses). ECC data

Cluster	**1** (n=231)	**2** (n=250)	**3** (n=93)	**4** (n=49)	**5** (n=14)	**6** (n=40)	**7** (n=23)	**8** (n=163)	**9** (n=189)	**10** (n=145)	**11** (n=471)	**Total** 1668
Number of families	115 (50%)	229 (92%)	80 (86%)	28 (57%)	5 (36%)	11 (28%)	8 (35%)	128 (79%)	174 (92%)	109 (75%)	212 (45%)	568 (66%)

The percentages vary widely by cluster in Table 31, and the largest percentages of families who met the reduced criteria were in clusters 2, 3 and 9. However, the main features of these clusters were child safeguarding issues (CPP, LAC and CIN), which was not a criterion; children in these clusters had low levels of school absence, exclusion and criminal offences anyway and therefore were likely to fall under the thresholds. Whilst these percentages provide some insight, they do not fully satisfy the Phase 1 criteria, given the missing data.

The Phase 2 guidelines were less specific and simply stated that families could be considered 'turned around' where there had been significant progress compared to their problems at the point of engagement. This definition was utilised for the following analysis, as it allowed consideration of all the available data. Analysis was performed to consider whether a family's circumstances had improved or worsened. At the most basic level, any families that had no further events after the start of intervention could be considered to have had an improvement in their circumstances (or, where they had no events before the start of intervention as well, their circumstances had remained the same but not worsened). And families that had a decrease in the number of different events they had might also be considered to have had some improvement (for example, if before the start of intervention, a family had school exclusion, school absence and CIN

events, but after had only school absence, this could be considered an improvement). In contrast, families who had an increase in the number of different events could be considered to have had a worsening of their circumstances (for example, where 'before' they had only school absence, but 'after' they had school absence and exclusion, this would be considered a worsening in circumstances).

For each cluster, Table 32 details the number (and percentage) of families who had no further events, fewer different types of events or more events after the start of intervention; it also includes a total for the whole group of families.

Table 32: Number (and percentage in parentheses) of families who had no further events, fewer events, or more events after the start of intervention. ECC data

Cluster	**1** (n=231)	**2** (n=250)	**3** (n=93)	**4** (n=49)	**5** (n=14)	**6** (n=40)	**7** (n=23)	**8** (n=163)	**9** (n=189)	**10** (n=145)	**11** (n=471)	**Total** 1668
Families with no further events	34 (15%)	51 (20%)	11 (12%)	5 (10%)	5 (36%)	1 (3%)	2 (9%)	21 (13%)	85 (45%)	11 (8%)	342 (73%)	568 (34%)
Families with fewer Events	80 (35%)	109 (44%)	43 (46%)	18 (37%)	2 (14%)	22 (55%)	6 (26%)	0	0	62 (43%)	0	342 (21%)
Families with more Events	51 (22%)	33 (13%)	10 (11%)	12 (24%)	1 (7%)	6 (15%)	11 (48%)	52 (32%)	42 (22%)	31 (21%)	129 (27%)	378 (23%)

Table 32 highlights that 34% of families overall had no further events after the start of intervention treatment. However, that figure includes the families from cluster 11 who had no events before intervention either. Overall, 14% of families who had events before intervention then had no more after intervention, which represented an improvement in their circumstances. There was wide variation between the clusters, with the clusters having between 1 and 85 families who had no further events (excluding cluster 11). Setting aside cluster 11, the cluster with the highest percentage of families with no further events was cluster 9 (45%). Families in cluster 9 had only CIN events 'before', and as has been previously noted, there was an overall reduction in the percentage of families with CIN events, so this may explain the high percentage of families from cluster 9 who no longer had events. In contrast, only 1 family (3%) from cluster 6 had no further events after the start of intervention. Cluster 6 represented families with a complex mix of issues, therefore it was perhaps unlikely that many of these families would have had no further events in the following year.

Where the percentage of families with fewer different events after the start of intervention was considered, all but clusters but 8, 9 and 11 had families who had fewer different events. Since families from clusters 8 and 9 only had one type of event before intervention (school absence and CIN events, respectively), they could not have 'fewer' events after; if there was a decrease, they had no events. Similarly, since families in cluster 11 had no events, it was not possible to have fewer events after intervention. Overall, just over a fifth (21%) of families had fewer different events after the start of intervention treatment, which could be considered an improvement in their circumstances. Over half (55%) of families from cluster 6 had fewer events.

All clusters had a proportion of families who had an increase in the different types of events they had following the start of intervention. Overall, 23% of families had an increase. Where families had an increase, it could be considered that their situations had worsened after the start of intervention. Cluster 7, in particular, had the highest percentage of families with an increase, with just under half (48%) of families. Cluster 5 had the lowest percentage (7%) of families.

Overall, considering Table 32, it is clear that there were improvements for some families after the start of intervention treatment when compared to before. 34% of families had no further events, although it must be considered that 60% of these families had no events before either. 21% of families had fewer events; therefore, considered together 55% of families had no, or fewer different events following the start of intervention, and could be considered to have had some improvement in their circumstances. The remaining 45% of families had either no change (50%) or an increase in the number of different events that they had.

However, this analysis did not consider the levels of individual events; that is, it considered, whether a family had an event (for example, school absence), but not the frequency of this (that is, it didn't compare how much school absence there was 'before' and 'after' the start of intervention). As a final analysis, and giving consideration to the Government criteria, the analysis already carried out in this section, and the information that was actually available in the database, two criteria for 'improvement' in a family's circumstances were considered. A 'strict' criteria, and a 'relaxed' (and perhaps more realistic) criteria were derived.

The ‘strict’ criteria considered ‘improvement’ to have occurred where a family had no further events after the start of intervention, or where they had events but had a reduction in the occurrence of all of those events. A reduction would be less school absence, fewer exclusions and fewer criminal offences (child and adult), plus no further CIN, CPP, LAC or NEET events (if a family had any of these to start with). These criteria were strict, as every single event that a family had before intervention would have to either have stopped or else had fewer occurrences of, for it to be considered that there was an improvement. Whilst it was possible to consider a reduction in the countable events (school absence, exclusion, and criminal offences), it was not possible to do this for CIN, CPP, LAC and NEET events as they were counted as a binary ‘yes’ or ‘no’; a reduction for them would be zero. As an example, if a family had school absence, school exclusion and CIN events before the start of intervention, an ‘improvement’ would be that they had less absence, fewer exclusions and no further CIN events; a decrease in only one or two of these events would not count.

The ‘relaxed’ criteria considered ‘improvement’ to occur where a family had no further events after the start of intervention, or where they had events but had a reduction in the occurrence of all of the countable events. That is, less school absence, fewer exclusions and fewer criminal offences (child and adult), if a family had these ‘before’. For the relaxed criteria, a family could still have CIN, CPP, LAC or NEET events ‘after’ if they had had them ‘before’, as long as there was a reduction in the other countable events. This was still fairly strict, as it required a decrease in all of the countable events, but allowed that existing safeguarding, or NEET issues could continue where there had been some other reduction. This is perhaps more in line with the Government guidelines, as it did not implicitly consider safeguarding issues in regard to a family being ‘turned around’.

These criteria were considered more appropriate (given the available data) than the attempted approximation of the Government guidelines. This is because, where the Phase 1 guidelines were attempted (Table 31), this considered whether school absence was under 15% and whether there were less than 3 exclusions, however, this meant that a family could have an increase in absence and exclusion, and as long as it was under those thresholds, still be considered to have had an improvement. Instead, the ‘strict’ and ‘relaxed’ criteria derived (and listed in Table 33) consider only actual reductions in events (rather than satisfying a threshold) and also consider the Phase 2 guidelines which

looked for sustained evidence of improvement (that is, the family have no increase in the different types of events that they have). The families who meet these derived criteria are not referred to as 'turned around' in this analysis, since it was not possible to fully consider the Government guidelines; rather, it is considered that they show some evidence of an 'improvement' in their circumstances in the year following the start of intervention treatment.

Table 33: Families who had some improvement after the start of intervention, using a combined approximation of the Government guidelines with consideration of the available ECC data, by cluster

Cluster	**1** (n=231)	**2** (n=250)	**3** (n=93)	**4** (n=49)	**5** (n=14)	**6** (n=40)	**7** (n=23)	**8** (n=163)	**9** (n=189)	**10** (n=145)	**11** (n=471)	**Total** 1668
Families satisfying strict criteria	51 (22%)	64 (26%)	12 (13%)	8 (16%)	7 (50%)	14 (35%)	4 (17%)	55 (34%)	85 (45%)	37 (26%)	342 (73%)	679 (41%)
Families satisfying relaxed criteria	58 (25%)	108 (43%)	26 (28%)	16 (33%)	7 (50%)	18 (45%)	4 (17%)	55 (34%)	85 (45%)	45 (31%)	342 (73%)	764 (46%)

Considering the 'strict' criteria, 41% of families overall had 'improvement' after the start of intervention. Cluster 11 had the highest percentage of families with improvement; this is because, as previously discussed, all families in cluster 11 had no events prior to intervention and many (73%) still had no events in the year following the start of intervention. Clusters 5 and 9 had the next highest percentages (50% and 45% respectively). Cluster 5 contained families who all had adults who had committed criminal offences before intervention, whereas cluster 9 contained families who all had only CIN events before. Cluster 7 had the lowest percentage of families (17%) who showed some improvement after intervention. However, the families in cluster 7 had a diverse mix of events before and many of them could not satisfy the strict criteria, which required improvement in all events after the start of intervention.

Where considering the 'relaxed' criteria, 46% of families overall had improvement after the start of intervention treatment. This was not a massive increase over the 'strict' criteria, but the slight relaxation of the rules allowed an extra 85 families (5%) to be considered as having had an improvement in their events after the start of intervention. Clusters 5, 7, 8, 9 and 11 had no change in percentages compared to the 'strict' criteria.

Table 34 compares 'improvement' (under the 'relaxed' criteria) with the prevalence of planned and unplanned endings. The left side of the table considers only those families that had an 'improvement' and details the percentage of first interventions that ended in

planned or unplanned endings, by cluster. The right side of the table considers the families that had no 'improvement' under the relaxed criteria (that is, the level of their events either stayed the same, or got worse after the start of intervention). The two groups are listed together to enable a direct comparison.

Table 34: Percentage of first interventions that ended in planned or unplanned endings, by families that had 'improvement' or not, and by cluster

Families with improvement				**Families without improvement**		
Cluster	**Planned Ending**	**Unplanned Ending**		**Planned Ending**	**Unplanned Ending**	**Cluster**
1 (n=58)	84%	12%		74%	20%	**1** (n=173)
2 (n=108)	84%	13%		78%	20%	**2** (n=142)
3 (n=26)	73%	27%		81%	16%	**3** (n=67)
4 (n=16)	81%	19%		70%	27%	**4** (n=33)
5 (n=7)	57%	42%		100%	0	**5** (n=7)
6 (n=18)	89%	11%		72%	23%	**6** (n=22)
7 (n=4)	50%	25%		84%	16%	**7** (n=19)
8 (n=55)	76%	16%		72%	22%	**8** (n=108)
9 (n=85)	72%	21%		76%	20%	**9** (n=104)
10 (n=45)	69%	24%		76%	19%	**10** (n=100)
11 (n=342)	74%	22%		75%	22%	**11** (n=129)
All families (n=764)	76%	20%		76%	20%	**All families** (n=904)

Superficially, it might seem logical to imagine that families who had 'improvement' would have planned endings for their first intervention treatment, however, this was not the case. Overall, just over three quarters of families (76%) who had an 'improvement' after the start of intervention had a first intervention that concluded as a planned ending; and this percentage was the same for the 'no improvement' group. For both groups, a fifth of first interventions (20%) had an unplanned ending. The fact that there was no difference overall between groups might suggest that the first intervention did not contribute to the family's 'improvement', however, it is important to consider that some families received more than one intervention treatment. And also, that a planned ending might simply mean that a family cooperated and met the requirements of their particular treatment plan, rather than that it indicated progress in terms of events. It is possible that planned and unplanned endings might be considered a measure of cooperation, or engagement, with the TF Programme, but that they (or at least the first intervention) may not provide an indication of whether a family showed 'improvement' in terms of the occurrence of events.

Considering both groups in Table 34, the percentages vary by cluster. For cluster 5, all families who showed 'no improvement' had planned endings to their first treatment; whereas just over half (57%) of the families that showed 'improvement' had planned endings. Whilst these percentages were more extreme than for the other clusters, and may reflect the small sample size of cluster 5, they do highlight the counter-intuitive relationship between planned endings and 'improvement' in the frequency of events that occurred.

Table 35 compares the two groups (families with and without 'improvement') with the percentage of families who received more than one type of intervention treatment, by cluster. Families received more than one type of treatment where their needs were complex, and the first intervention treatment was perhaps not enough to help on its own.

Table 35: Percentage of families who received more than one intervention, for families with and without 'improvement' and by cluster assignment. ECC data

Families with improvement			**Families without improvement**	
Cluster	**Percentage of families who received further treatment**		**Percentage of families who received further treatment**	**Cluster**
1 (n=58)	28%		42%	**1** (n=173)
2 (n=108)	39%		50%	**2** (n=142)
3 (n=26)	35%		46%	**3** (n=67)
4 (n=16)	19%		42%	**4** (n=33)
5 (n=7)	29%		57%	**5** (n=7)
6 (n=18)	50%		41%	**6** (n=22)
7 (n=4)	25%		47%	**7** (n=19)
8 (n=55)	24%		42%	**8** (n=108)
9 (n=85)	26%		53%	**9** (n=104)
10 (n=45)	31%		58%	**10** (n=100)
11 (n=342)	28%		46%	**11** (n=129)
All families (n=764)	30%		47%	**All families** (n=904)

The table highlights that, overall (and for almost all clusters) higher percentages of families who showed no improvement had further intervention treatment. This indicates that families with more complex needs (who did not show improvement after a year) were more likely to receive further interventions, which would seem logical. It also indicates that, whilst the success (or not) of a first intervention may not aid in indicating quantifiable improvement for a family, whether a family had further intervention treatment was more closely linked to 'improvement'.

It should be considered that not all families who did not satisfy the criteria for 'improvement' necessarily had any escalation in their issues, they just did not show 'improvement'. Therefore, the data was also analysed in order to consider only families who had an escalation in their issues. That is, they had an increase in school absence, and more school exclusions and criminal offences (if they had them 'before'), and had occurrences of CIN, CPP, LAC or NEET events (where they had not had them 'before'). This identified that just over a quarter of families (26%) had more, or greater occurrences, of events in the year following the start of intervention.

7.3.5.1 Summary

This section considered the Government's guidelines on what constituted a family being 'turned around'. There was no definitive information regarding this within the database, and since the data that was available could not satisfy the Government guidelines, it was not possible to determine whether any of the families might have been 'turned around' in the year following their introduction to the TF programme. However, in consideration of these guidelines, and given the available data, new criteria were derived that attempted to indicate where a family had some improvement. Using the 'relaxed' criteria, 'improvement' was classified as having no further events after the start of intervention, or having a reduction in all of the countable events (school absence, exclusion and criminal offences, if a family had them before), and no increase in CIN, CPP, LAC or NEET events. Overall, just under half of families (46%) had some 'improvement' in the year following their introduction to the TF programme, and the percentages varied quite widely between clusters.

Where consideration was given to how a family's first intervention treatment ended (planned or unplanned endings), it seemed that the success, or not, of the first intervention may have had little impact upon whether a family showed improvement in terms of the events that occurred. However, a greater percentage of families who did not show improvement received further intervention treatment (they received more than one type), which cautiously indicates that the lack of improvement was recognised in some cases and families received further treatment in an attempt to rectify this.

It should be considered that it was not possible to identify a comparison group from the data so that a more thorough analysis could be performed in order to determine whether improvement might be directly attributed to the TF programme. This was because it was

very difficult to identify families with similar needs that were not already receiving treatment. The ECC also had this difficulty and felt that generally most families who had needs qualifying them for the TF programme had already been identified and were receiving some form of help. Those few families not receiving treatment, and that might form a comparison group, in general, had fewer needs. Another problem, pertaining to this particular analysis, is the notion of having some sort of start date (first intervention date) around which to compare events; there may be no equivalent date for comparison families. However, whilst difficult to identify, further consideration of whether it is possible to identify any kind of useful comparison group (perhaps if more data were to become available in future), could be a useful avenue for future research.

7.3.6 Detailed Summary of clusters following the start of intervention

Collecting together the data from the previous sections (school absence, events in the year following intervention, and the consideration of the outcome) and the cluster data from the previous chapter, there follows a detailed summary of each of the clusters in the year following their introduction to the TF programme, in comparison to the year before. This is in order to consider any changes that might have occurred for families after joining the TF programme

For each cluster, a Slopegraph was plotted to highlight the changes. The Slopegraph plots the percentage of families with each of the eights event (that were clustered upon) in the year before the start of intervention, and also the percentage in the year after the start of intervention. For the 'before' data, all 2155 families were included, and for the 'after' data all 1668 families who had a years' worth of 'after' data available were included. The slopegraph provides a visualisation of the changing trends within each cluster.

First, to provide contrast, Figure 70 plots the slopegraph for all families (that is, not on the cluster-level). Aside from the decrease in the occurrence of CIN events (which has been previously discussed), the plot indicates that there was very little change. If anything, there was an overall slight increase in the occurrence of events for the families where comparing 'before' and 'after'. However, as will be highlighted in each cluster summary, on the cluster-level there were significant changes.

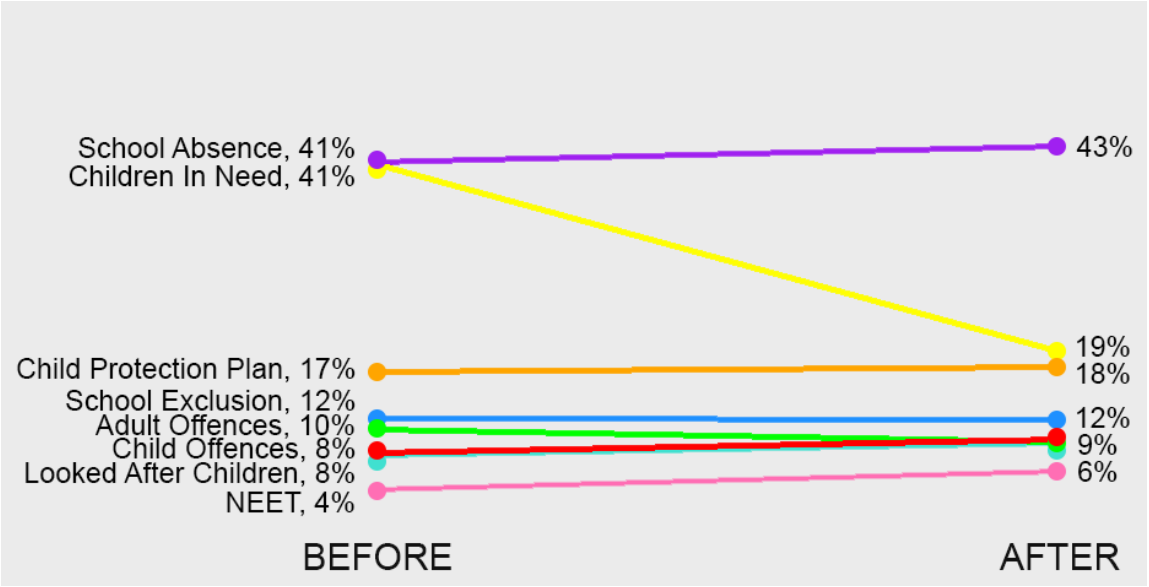


Figure 70: Percentage of families with events in the years before and after the start of intervention, for all families

7.3.6.1 Cluster 1: School exclusion and criminal offences

Cluster 1 contained 291 families, 231 families (79%) had available data for the 'after' analysis.

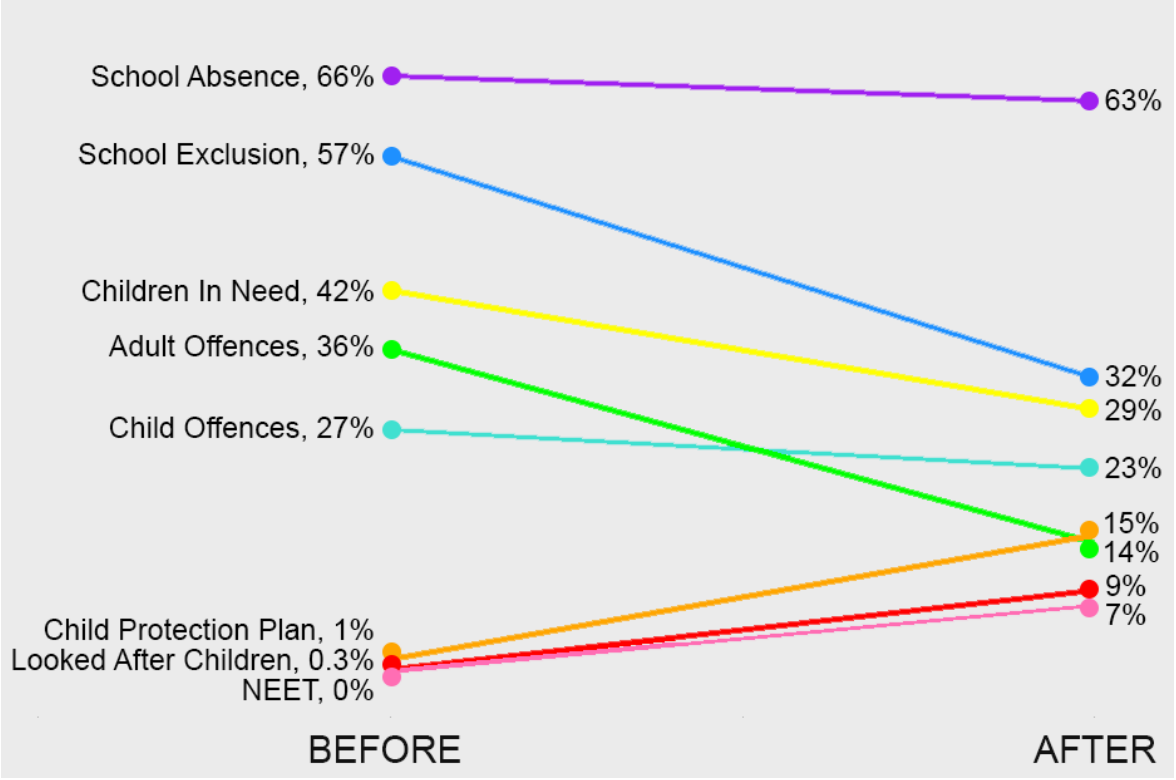


Figure 71: Percentage of families with events in the years before and after the start of intervention for cluster 1

As Figure 71 highlights, the primary features of this cluster, school exclusion and criminal offences, decreased in the year following the start of intervention. Indeed, the five most populous events (school absence, exclusion, CIN and criminal offences) all decreased.

The percentage of families with school exclusion had the largest decrease (from 57% to 32%) and whilst 'before' two thirds (66%) of all families with school exclusion were contained in this cluster, this decreased to 39% 'after'. There was a large decrease in the percentage of families with adults who committed criminal offences (from 36% to 14%), however only a small decrease (of 4%) for criminal offences committed by children.

Whilst the percentage of families with CIN events dropped (from 42% to 29%), there was an increase in families with more serious child safeguarding events after the start of intervention (with CPPs increasing from 1% to 15%, and LACs from 0.3% to 9%).

Whilst there was a small decrease in the percentage of families with school absence, the amount of school absence per family increased slightly. The average levels of unauthorised absence per family increased (from an average of 3.9% unauthorised school sessions per family, to 7.3%), and the percentage of families that had absence greater than 15% increased (from 7% to 13%). There was also an increase in families with NEET members (7% of families compared to none before).

It was notable that 15% of families in this cluster had no events at all in the year following the first intervention date, which would indicate a positive change for these families. And where the cluster rules from the 'before' analysis were applied, just under a third of families (31%) would have remained in cluster 1, that is, their circumstances remained the same. For those that changed cluster, 15% would be assigned to cluster 11 (no events), 13% to cluster 2 (Child Protection Plans) and 13% to cluster 8 (just school absence); the rest were split over the remaining clusters.

Three quarters (75%) of first interventions ended in a planned ending, 18% of families had an unplanned ending and 7% were still continuing a year later. A quarter of families had an improvement in their circumstances after the start of intervention according to the 'relaxed' criteria; this was low in comparison to the other clusters, only cluster 7 had a lower percentage. Just over a third (37%) of families overall received further intervention treatment. Overall, 42% of families who had shown no improvement received further intervention treatment, compared to 28% of families who had shown improvement.

7.3.6.2 Cluster 2: Child Protection

Cluster 2 contained 335 families, 250 families (75%) had available data for the 'after' analysis.

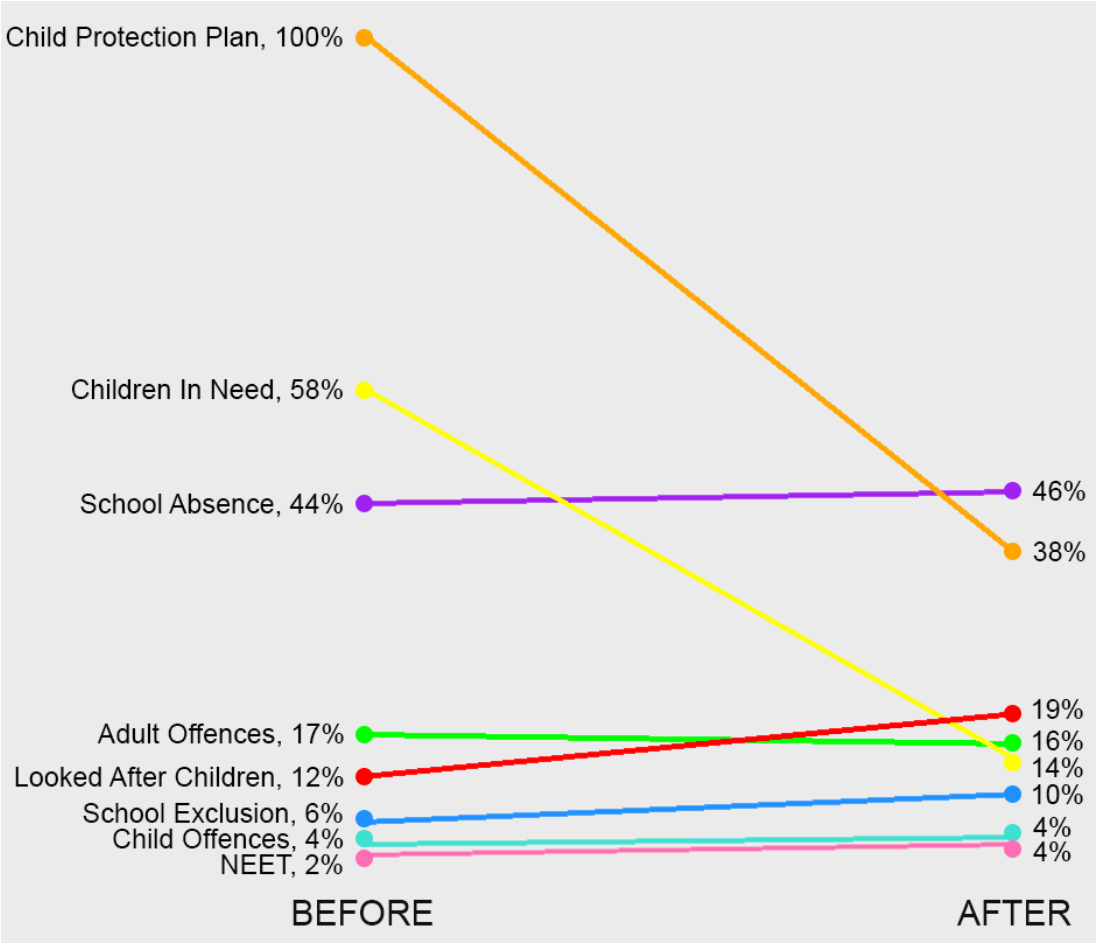


Figure 72: Percentage of families with events in the years before and after the start of intervention for cluster 2

Figure 72 highlights the biggest change in the 'after' analysis was a large decrease in the percentage of families with CPP and CIN events. The primary feature of this cluster, Child Protection Plans, showed a large decrease, from all families having them 'before', to just over a third of families (38%) 'after'. Whereas 'before' almost all of the families with CPP events (92%) were contained in this cluster, this decreased to only a third (33%) following the start of intervention. The percentage of families with CIN events also decreased. Overall, this represented a significant reduction in the percentage of families with child safeguarding issues. However, there was an increase (from 12% to 19%) in the percentage of families with LAC events, indicating that a small number of family's child safeguarding needs escalated after the start of intervention.

Overall, the percentage of families with the other events showed little change. School absence increased slightly (from 44% to 46%), but the average levels of unauthorised

absence decreased (from 3.7% per family on average to 2.6%, and from 15% of families with over 15% absence to 4%). The percentage of families with school exclusion and those with NEET members both increased slightly (by 4% and 2% respectively). The percentage of families with criminal offences stayed almost unchanged, for both adult and child offences.

It was notable that 20% of families in this cluster had no events at all in the year following the first intervention date, which would seem to indicate a positive change for these families. And where the cluster rules from the 'before' analysis were applied, just over a third of families (36%) would have remained in cluster 2, that is, their circumstances remained the same. For those that changed cluster, 20% would be assigned to cluster 11 (no events), 13% to cluster 3 (Looked after Child events) and 10% to cluster 1; the rest were split over the remaining clusters.

80% of first interventions ended in a planned ending (which was high in comparison to the other clusters), 17% had an unplanned ending and 4% were still continuing a year later. 43% of families had an improvement in their circumstances after the start of intervention according to the 'relaxed' criteria. 41% of families overall received further intervention treatment, of these 62% had shown no improvement. Overall, where families showed no sign of improvement in the year following the start of intervention, half received further treatment and half did not.

7.3.6.3 Cluster 3: Looked After Children

Cluster 3 contained 115 families, 93 families (81%) were included in the 'after' analysis.

The primary feature of this cluster was that all families had Looked After Children events; this decreased to 42% of families following the start of intervention. Figure 73 highlights that there was also a notable decrease in the percentage of families with CIN events (from 63% to 16%), however there was a large increase in the percentage of families with Child Protection Plans (from 2% to 32%). This suggests that, overall, child safeguarding concerns had been downgraded for many of the families; 58% of families no longer had children in care events, but instead just under half of them had Child Protections Plans. There was a large decrease in the percentage of families who had changed address at least once (from 73% 'before' to 34% 'after'), this might reflect that since there were fewer families with children in care, there were therefore fewer address changes (as family members were not moving around as much).

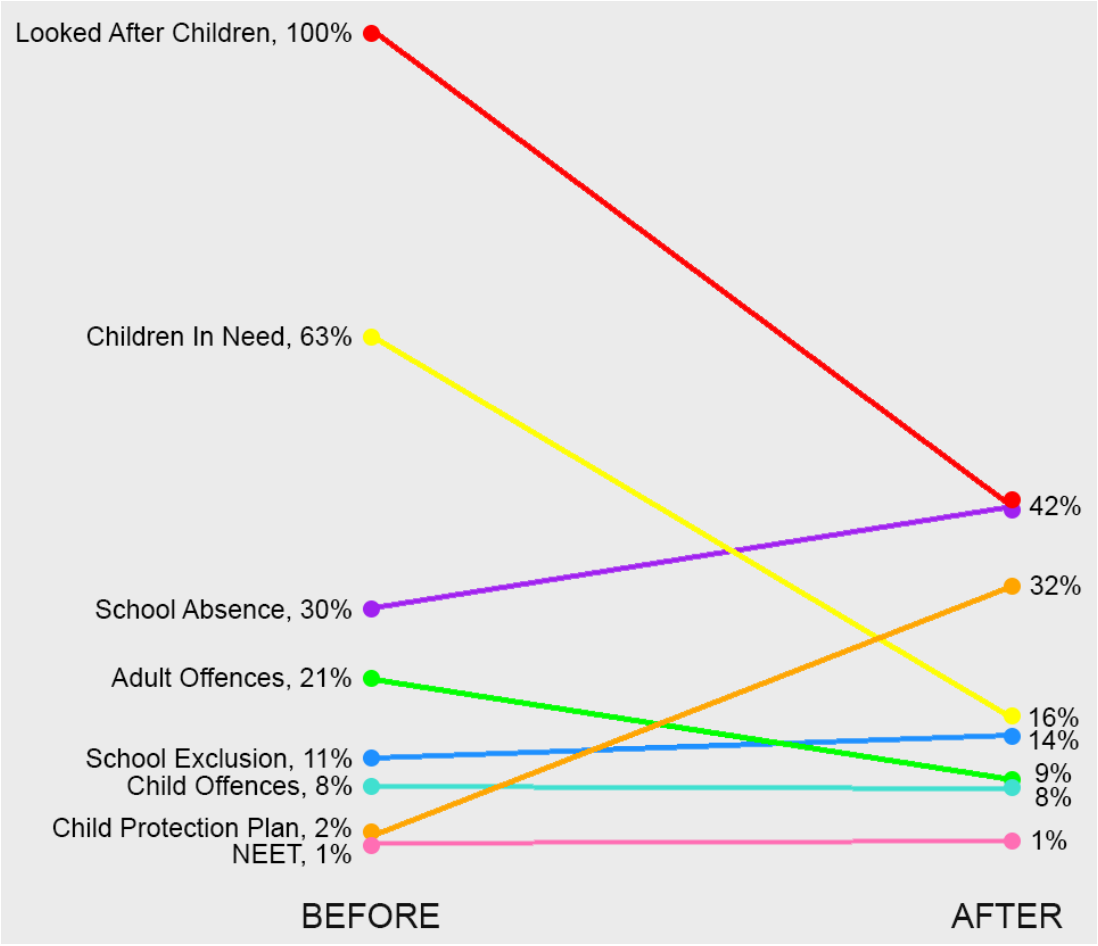


Figure 73: Percentage of families with events in the years before and after the start of intervention for cluster 3

There was a decrease in criminal offences committed by adults (from 21% of families to 9%), although offences committed by children remained the same (8% of families). School absence and exclusion both increased (absence from 30% to 42% of families, and exclusion slightly from 11% to 14%).

It was notable that 12% of families in this cluster had no events at all in the year following the first intervention date, which would seem to indicate a positive change for these families. And where the cluster rules from the 'before' analysis were applied, just over a third of families (37%) would have remained in cluster 3, that is, their circumstances remained the same. For those that changed cluster, 32% would be assigned to cluster 2 (Child Protection Plans), and 12% to cluster 11 (no events); the rest were split over the remaining clusters except for clusters 4 and 7.

81% of first interventions resulted in a planned ending (this was the highest percentage over all clusters), 17% had an unplanned ending and 2% were still continuing a year later. 28% of families had an improvement in their circumstances after the start of intervention according to the 'relaxed' criteria, which was fairly low compared to the other clusters

(only clusters 1 and 7 had lower percentages). Just under half (44%) of families overall received further intervention treatment, of these 78% had shown no improvement. Overall, of the families who showed no sign of improvement (72%) in the year following the start of intervention, 46% received further treatment. Where families had shown signs of improvement, 35% received further treatment.

7.3.6.4 Cluster 4: NEETs

Cluster 4 contained 61 families, 49 families (80%) had available data for the 'after' analysis.

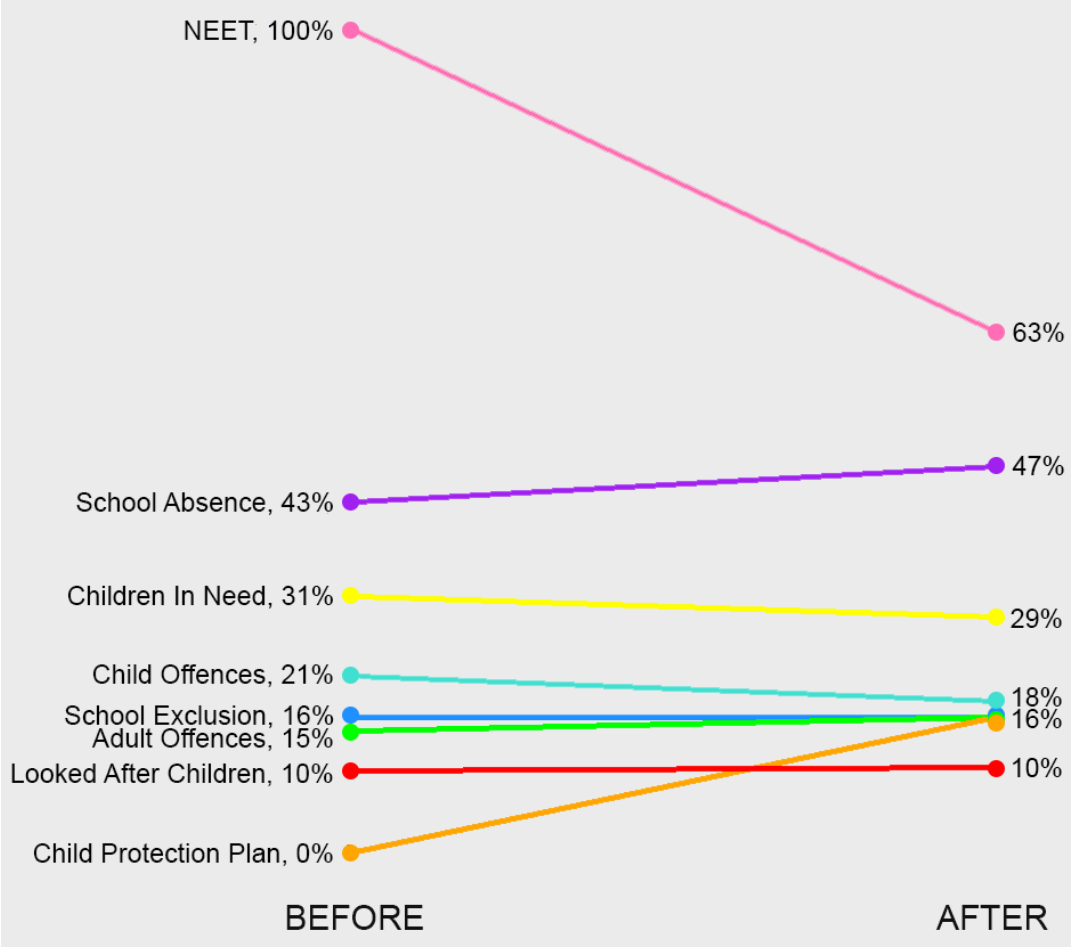


Figure 74: Percentage of families with events in the years before and after the start of intervention for cluster 4

The primary feature of this cluster was that all families had members who were Not in Employment, Education or Training (NEET), and this decreased to just under two thirds (63%) of families a year later. Figure 74 indicates that there was little change in the prevalence of most of the other events, however there was a notable increase in the percentage of families with Child Protection Plans (from zero to 16%). The percentage of families receiving DWP benefits increased from 57% 'before' to 78% 'after'. Whilst, as previously mentioned, there was an overall increase in the percentages of families

receiving DWP benefits for all clusters, this was the largest increase of all the clusters; it is not clear why.

10% of families in this cluster had no events at all in the year following the first intervention date, which would seem to indicate a positive change for these families. And where the cluster rules from the 'before' analysis were applied, just under half of families (47%) would have remained in cluster 4, that is, their circumstances remained the same. For those that changed cluster, 12% would be assigned to cluster 2 (Child Protection Plans), and 10% to cluster 11 (no events); the rest were split over the remaining clusters except for clusters 5, 6 and 10.

75% of first interventions resulted in a planned ending, 21% had an unplanned ending and 3% were still continuing a year later. A third (33%) of families had an improvement in their circumstances after the start of intervention according to the 'relaxed' criteria. Just over a third (35%) of families overall received further intervention treatment. Overall, of the families who showed no sign of improvement (67%) in the year following the start of intervention, 42% received further treatment. Where families had shown signs of improvement, 19% received further treatment.

7.3.6.5 Cluster 5: Adult Criminal Offences

Cluster 5 contained 21 families, 14 families (67%) had available data for the 'after' analysis.

The primary feature of this cluster was that all families had at least one adult who had committed a criminal offence, this decreased to 36% of families in the 'after' analysis. Figure 75 highlights that there was also a small reduction in the percentage of families with criminal offences committed by children (from 10% to 7%), and in families with school exclusions (from 24% to 21%).

A year later no family had a Child Protection Plan (decreased from 5%), however the percentage of families with CIN events had increased (from 10% to 21% of families). The increase in families with CIN events was unusual, as most of the clusters had a decrease. This represented an increase in lower-level child safeguarding events in the year following the start of intervention, but a decrease in the more serious safeguarding issues (there were no families with CPP or LAC events 'after'). Families in this cluster had a large increase in the percentage of drug/alcohol events (from zero 'before' to 14% 'after'), no

other cluster had an increase of more than a couple of per cent. However, there was a large decrease in the percentage of events classed as domestic abuse (from 33% 'before' to 14% 'after').

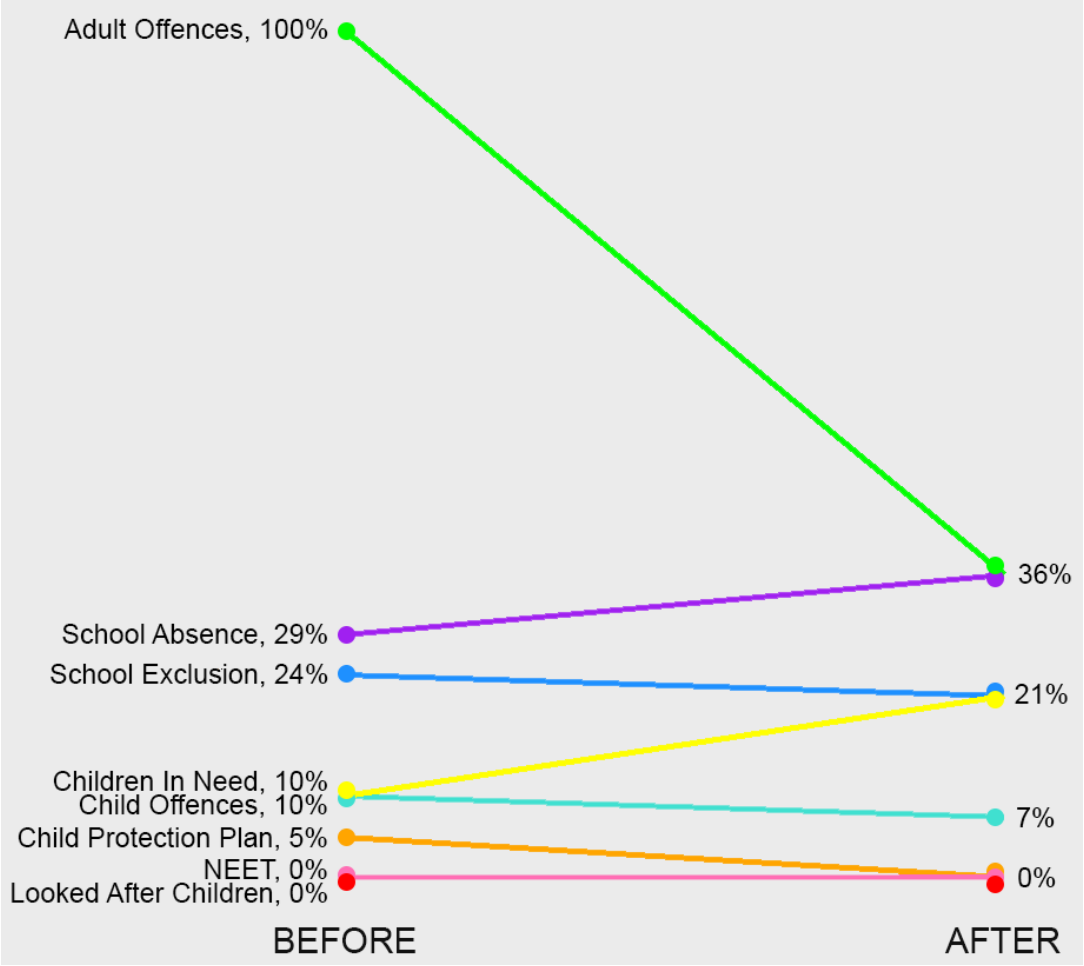


Figure 75: Percentage of families with events in the years before and after the start of intervention for cluster 5

It was notable that just over a third (36%) of families in this cluster had no events at all in the year following the first intervention date, which would seem to indicate a positive change for these families. And where the cluster rules from the 'before' analysis were applied, just over a fifth of families (21%) would have remained in cluster 5, that is, their circumstances remained the same. For those that changed cluster, 36% would be assigned to cluster 11 (no events), and 29% to cluster 1; the rest were split over the clusters 8 and 9.

57% of first interventions ended in a planned ending, and 43% had an unplanned ending; there were no continuing interventions. This was the highest proportion of unplanned endings across all of the clusters (cluster 8, with 24% unplanned endings was the next highest). However, despite this, half of the families (50%) had an improvement in their

circumstances after the start of intervention according to the 'relaxed' criteria – this was a higher percentage than for all clusters but cluster 11 (whose families had no events). However, it is important to consider the small size of this cluster when considering the data.

Just under half (48%) of families overall received further intervention treatment, of these 67% had showed no sign of improvement in the year following the start of intervention. Overall, of the families who showed no sign of improvement (50%) in the year following the start of intervention, 57% received further treatment. Where families had shown signs of improvement, 29% received further treatment.

7.3.6.6 Cluster 6: High Levels of School Absence

Cluster 6 contained 54 families, 40 families (74%) had available data for the 'after' analysis.

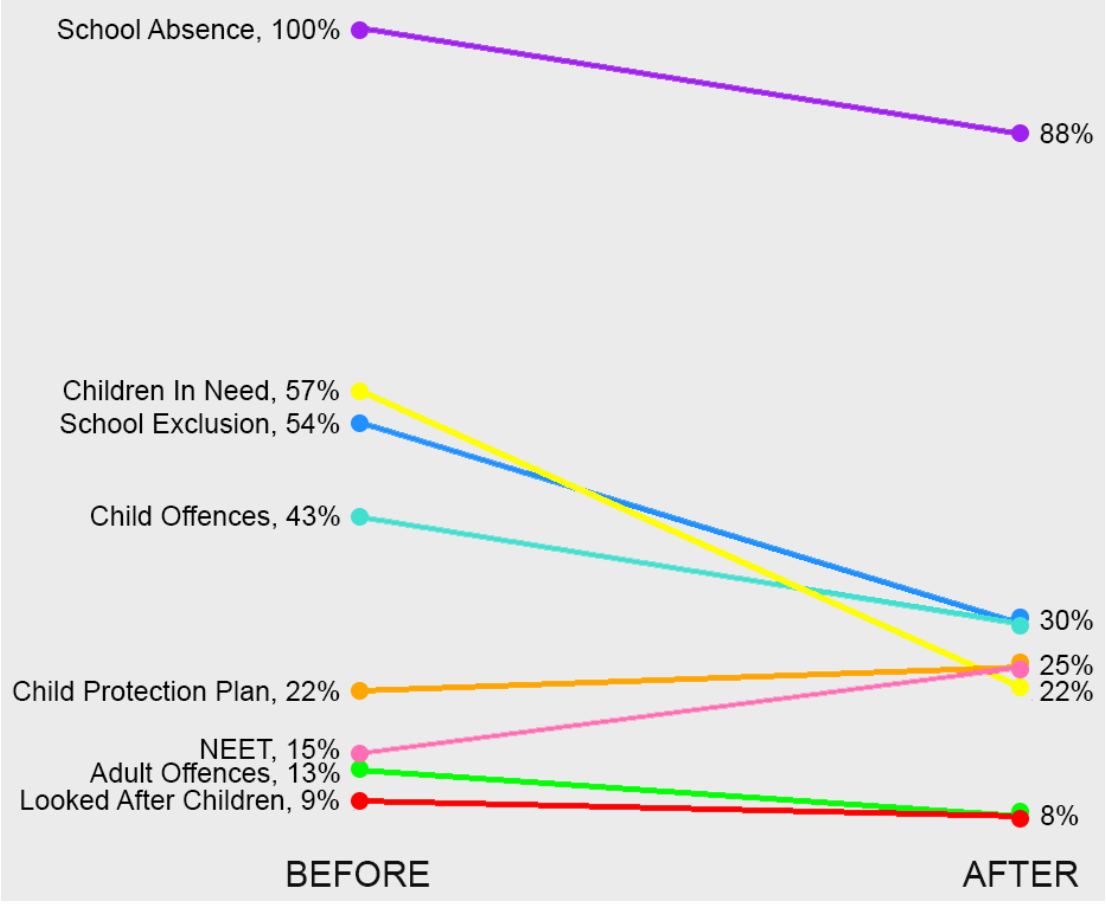


Figure 76: Percentage of families with events in the years before and after the start of intervention for cluster 6

This cluster's primary feature was that all families had high levels of school absence. In the year following the start of intervention, the percentage of families with school absence decreased from 100% to 88%, a much smaller decrease overall than was

recorded with the primary characteristics of the other clusters. However, where previously 98% of families had on average school absence that was greater than 15%, this decreased to 55% of families of families in the 'after' analysis. Therefore, although most families still had unauthorised school absence, the average amount for each family was less (overall it decreased from 38.7% to 21.9%). Figure 76 highlighted that there was also a reduction in the percentage of families with school exclusions (from 54% to 30%).

Considering child safeguarding, the percentage of families with CIN events decreased (from 57% to 22%), LAC events decreased slightly (from 9% to 8%) and CPP events increased a little (from 22% to 25%). There was therefore a decrease in lower-level safeguarding events, but the percentage of higher level events remained similar.

There were decreases in the percentage of families with criminal offences, for both those committed by adults and by children. Perhaps the most notable increase 'after' was the percentage of families with a NEET member (from 15% to 25% of families).

This was a diverse cluster, with a variety of events, and remained so a year later. Only 3% of families had no events at all in the year following the first intervention date, which was a lower percentage than for any of the other clusters. Where the cluster rules from the 'before' analysis were applied, only 8% of families would have remained in cluster 6, that is, their circumstances remained the same. For those that changed cluster, 28% would be assigned to cluster 8 (just school absence), and 18% to cluster 2 (CPPs); the rest were split over the clusters except for 5 and 9.

70% of first interventions resulted in a planned ending, and 19% had an unplanned ending; there were 7% of interventions continuing a year later. 39% of families overall received further intervention treatment. Overall, 45% of families had an improvement in their circumstances after the start of intervention according to the 'relaxed' criteria, and half of these received further intervention treatment. Of the families who showed no sign of improvement (55%) in the year following the start of intervention, 41% received further treatment. This was the only cluster where the percentage of families who received more than one intervention treatment was greater for the families who showed improvement, compared to those who did not show improvement.

7.3.6.7 Cluster 7: Child Criminal Offences

Cluster 7 contained 25 families, 23 families (92%) had available data for the 'after' analysis.

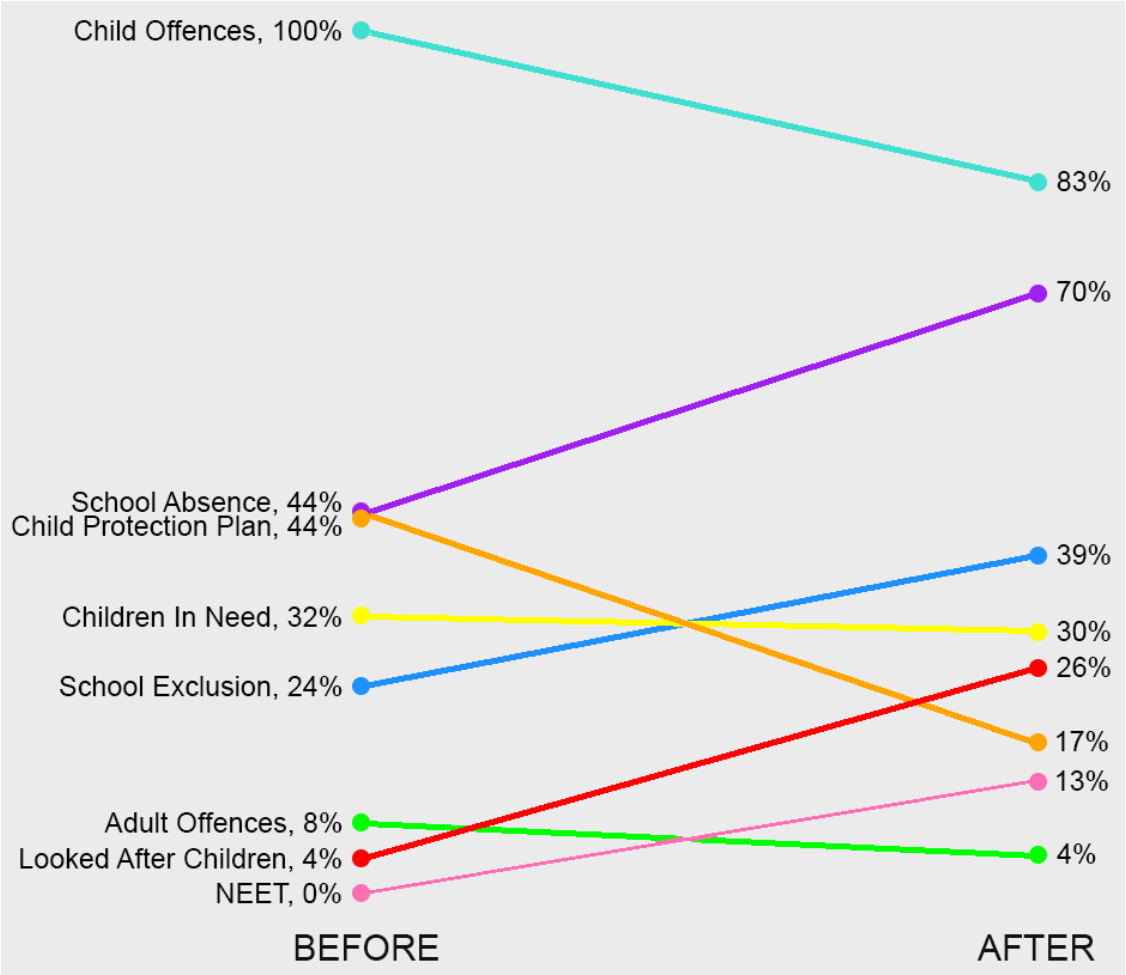


Figure 77: Percentage of families with events in the years before and after the start of intervention for cluster 7

The primary feature of this cluster was that it contained families who all had at least one child who had committed criminal offences. In the year following the start of intervention, the percentage of families with criminal offences committed by children decreased from 100% to 83%. There was also a small decrease in the percentage of families with criminal offences committed by adults (from 8% to 4%).

Perhaps the most notable change in Figure 77 was that the percentage of families with school absence and school exclusion increased. In the year before intervention 44% of families had school absence, this increased to 70% following the start of intervention; and families with school exclusion increased from 24% to 39%. There was also an increase in the percentage of families with NEET members.

The percentage of families with CIN events stayed almost the same (from 32% to 30%), and there was a decrease in families with Child Protection Plans (from 44% of families to 17%). However, the percentage of families with Looked After Children events increased from 4% to 22%, indicating an increase in serious child safeguarding issues.

9% of families had no events at all in the year following the first intervention date, which indicated a positive change for these families. Where the cluster rules from the 'before' analysis were applied, only 26% of families would have remained in cluster 7, that is, their circumstances remained the same. For those that changed cluster, 30% would be assigned to cluster 1, and 22% to cluster 3 (LAC events); the rest were split over clusters 2, 5, 6, 11.

72% of interventions ended in planned ending, 20% had an unplanned ending and 8% were ongoing a year later. Whilst only a fifth of families received AO treatment, all of the unplanned endings occurred for families receiving AO treatment. 40% of families received more than one intervention treatment. Overall, 17% of families had an improvement in their circumstances after the start of intervention according to the 'relaxed' criteria; this was the lowest percentage of all clusters. However, the slopegraph highlights that the overall prevalence of many of the events increased 'after', therefore it is perhaps unsurprising that so few families showed signs of improvement. Of the families who showed no sign of improvement (83%) in the year following the start of intervention, 47% received further intervention treatment.

7.3.6.8 Cluster 8: School Absence Only

Cluster 8 contained 223 families, 163 families (73%) had available data for the 'after' analysis.

This cluster contained families who all had school absence, but no other events in the year prior to intervention. However, there were a small percentage of families with pre-existing CPPs (12%) and LAC events (2%). Figure 78 highlights that the percentage of families with any school absence decreased from 100% to 79%. However, the levels of school absence remained the same (6% unauthorised sessions per family on average, and 12% of families had more than 15% unauthorised sessions).

The percentage of families with the other events increased (which was perhaps likely given they were zero prior to intervention), however the percentage of families with Child

Protection Plans decreased slightly (from 12% with pre-existing CPPs, to 8% a year after the start of intervention), although in the plot this is represented as an increase as pre-existing events were not counted. The greatest increase was for CIN events, with 22% of families having them 'after'. 9% of families had criminal offences that were committed by a child, whilst 4% had offences committed by adults. 7% of families had school exclusion 'after'. The percentage of families with LAC events and families with NEET members increased slightly (by 2%).

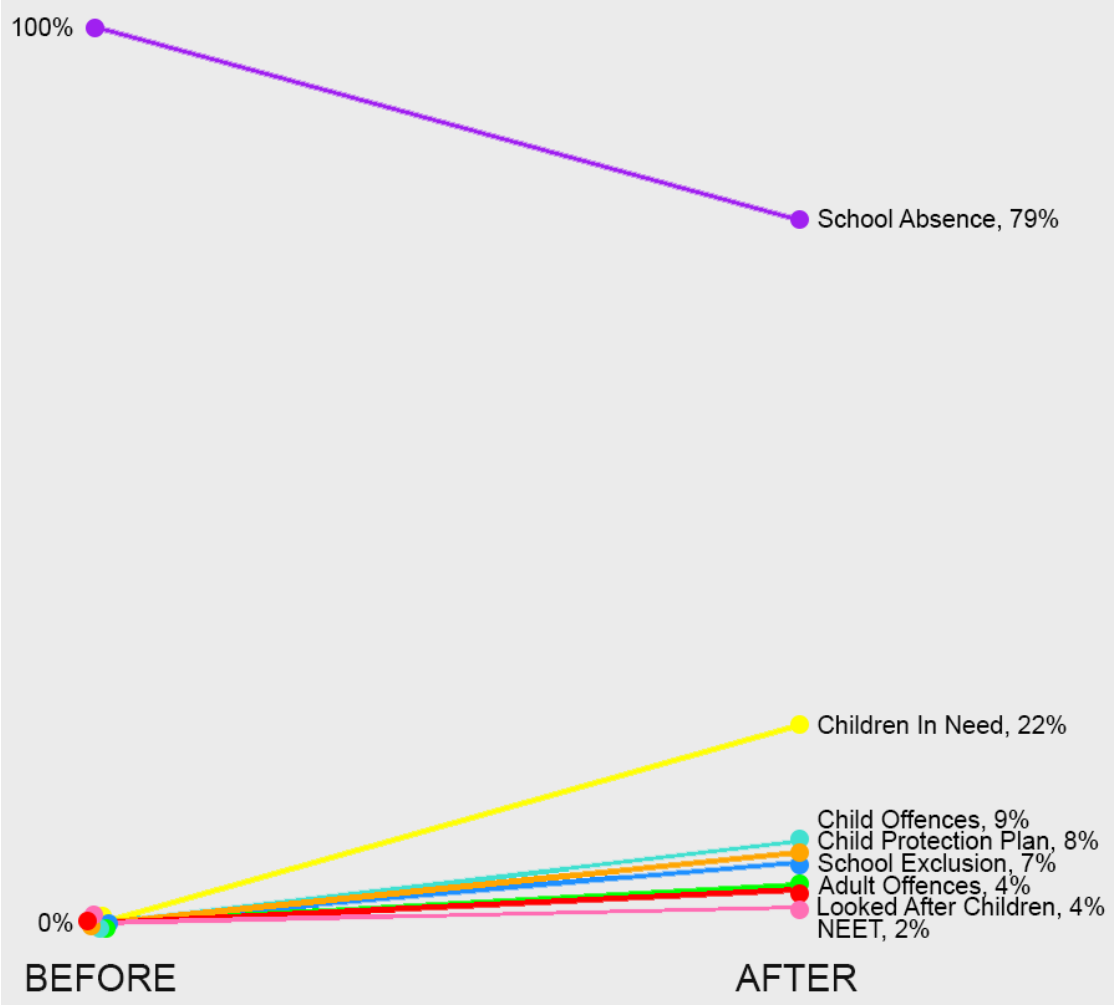


Figure 78: Percentage of families with events in the years before and after the start of intervention for cluster 8

13% of families had no events at all in the year following the first intervention date, which indicated a positive change for these families. Where the cluster rules from the 'before' analysis were applied, just under half (49%) of families would have remained in cluster 8, that is, their circumstances remained the same, and they only had school absence in the year following intervention. For those that changed cluster, 13% would be assigned to cluster 11 (no events), and 12% to cluster 10 (school absence and CIN events only); the rest were split over the other clusters.

69% of first interventions resulted in a planned ending (which was the second lowest percentage of planned endings compared to the other clusters), 24% had an unplanned ending and 7% were ongoing a year later. Just under a third (32%) of families received further intervention treatment, this was the second lowest percentage over all clusters. Overall, 34% of families had an improvement in their circumstances after the start of intervention according to the 'relaxed' criteria, and of these 24% received further intervention treatment. Of the families that showed no improvement (66%), 42% received further intervention treatment.

7.3.6.9 Cluster 9: Children in Need only

Cluster 9 contained 243 families, 189 families (78%) had available data for the 'after' analysis

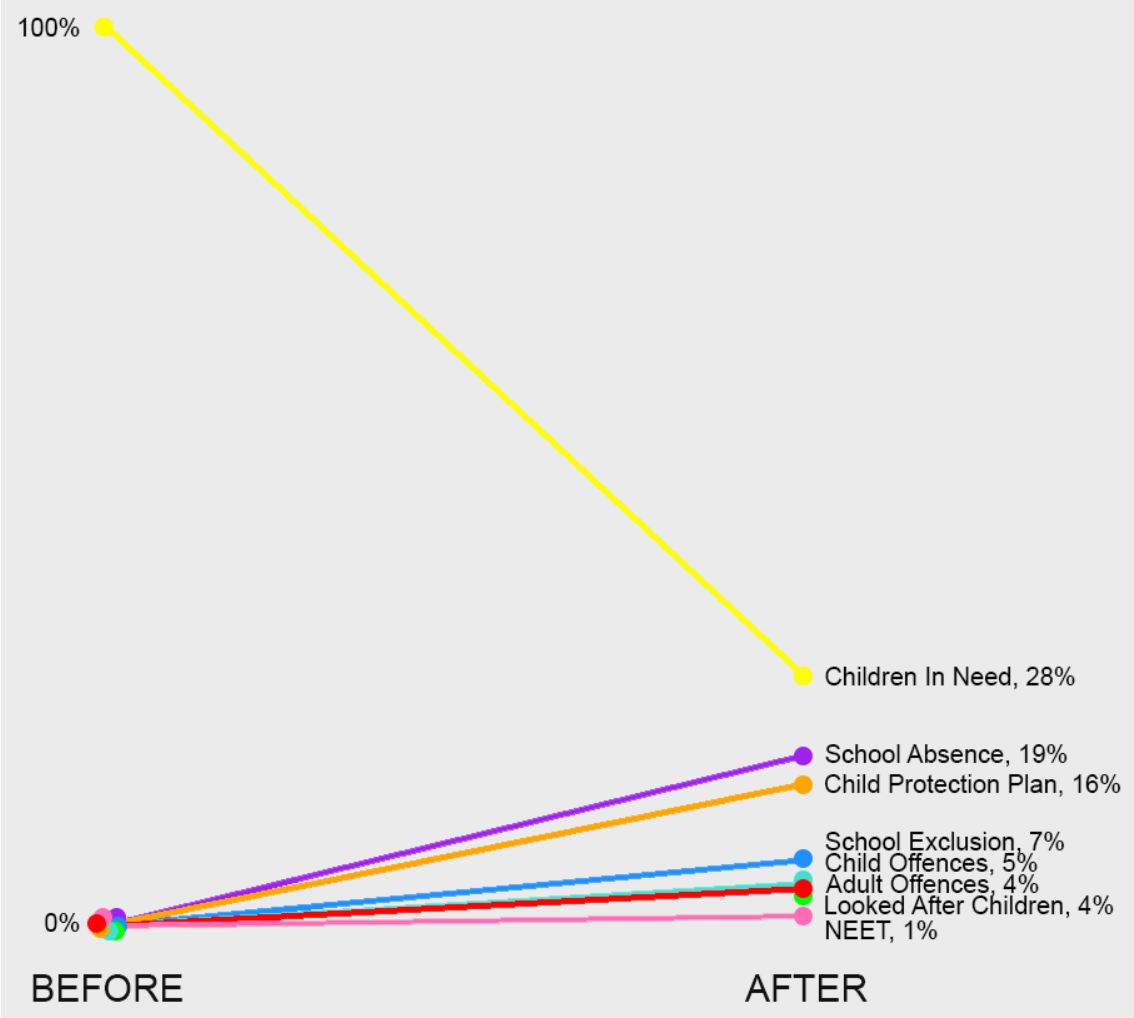


Figure 79: Percentage of families with events in the years before and after the start of intervention for cluster 9

This cluster contained families who all had at least one CIN event, but no other events in the year prior to intervention. However, there were a small percentage of families with

pre-existing CPPs (1%) and members who were NEET (1%). The percentage of families with CIN events decreased from 100% to 28%, as shown in Figure 79. However, there was an increase in families with Child Protection Plans, (to 16%) and a small increase in LAC events, suggesting that although there were far fewer families with lower-level child safeguarding events, there was a small group who changed to having higher-level events.

There was an increase in school related issues, just under a fifth (19%) of families had school absence in the year following the start of intervention, and 7% had school exclusions. There were small increases in the percentage of families with criminal offences committed by children (5%), and adults (4%).

45% of families had no events at all in the year following the first intervention date, which indicated a positive change for these families. This was the largest percentage of families (considering all clusters) who had no further events following the start of intervention. Where the cluster rules from the 'before' analysis were applied, 15% of families would have remained in cluster 9, that is, their circumstances remained the same and they only had CIN events in the year following intervention. For those that changed cluster, 45% would be assigned to cluster 11 (no events), and 16% to cluster 2 (CPP); the rest were split over the other clusters, except cluster 6.

75% of interventions resulted in a planned ending, 19% had an unplanned ending and 6% were ongoing a year later. 40% of families received further intervention treatment. Overall, 45% of families had an improvement in their circumstances after the start of intervention according to the 'relaxed' criteria, and this was the second highest percentage across all clusters; of these 26% received further intervention treatment. Of the families that showed no improvement (55%), 53% received further intervention treatment.

7.3.6.10 Cluster 10: School Absence and CIN

Cluster 10 contained 182 families, 145 families (80%) were included in the 'after' analysis.

This cluster contained families who all had unauthorised school absence and at least one CIN event, but no other events in the year prior to intervention. However, there were a small percentage of families (3%) that had pre-existing Child Protection Plans. As Figure 80 highlights, the percentage of families with CIN events decreased (from 100% to 22%), however the percentage with school absence decreased by a much smaller amount (from

100% to 87%). The amount of school absence that each family had remained the almost the same (10.6% average unauthorised sessions per family 'before' to 10.1% 'after').

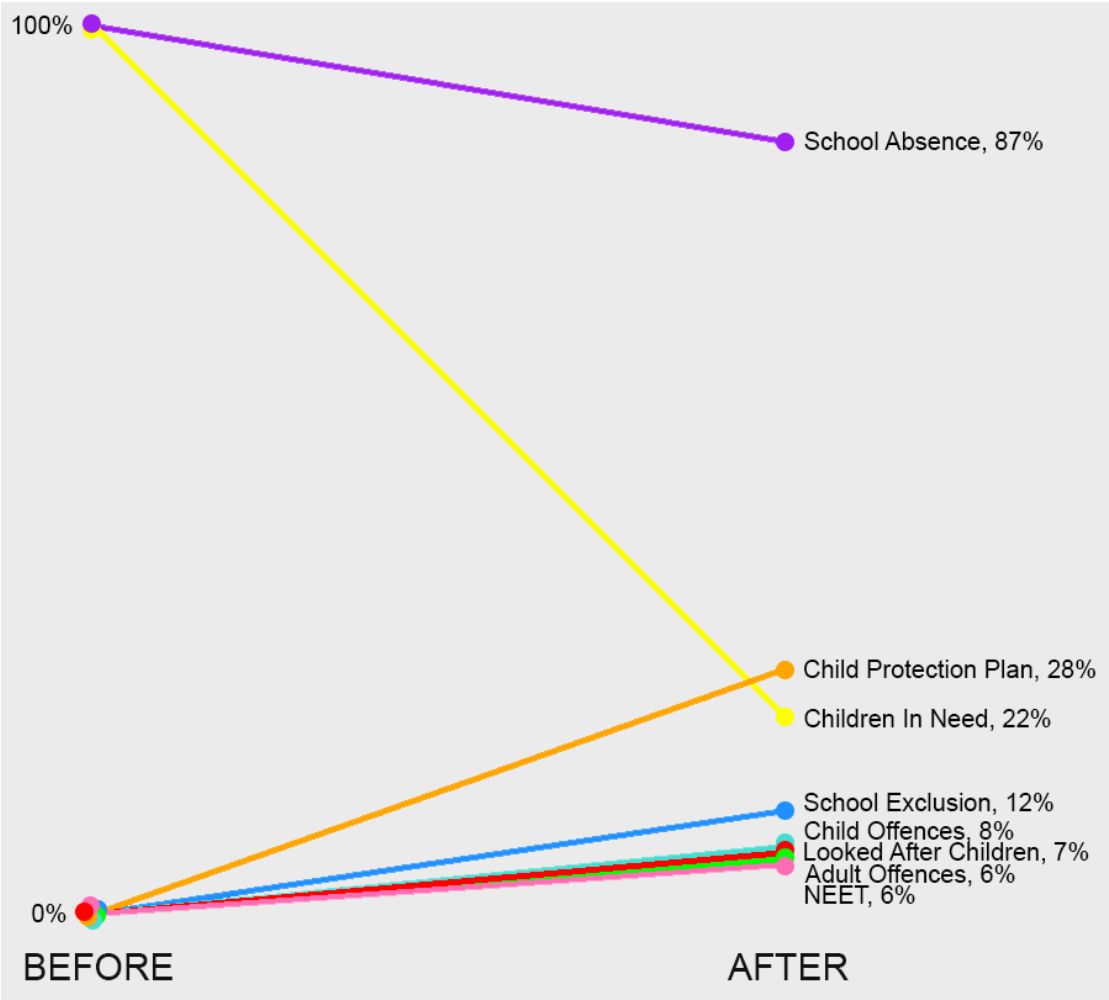


Figure 80: Percentage of families with events in the years before and after the start of intervention for cluster 10

The largest increase in the occurrence of a particular event was for families having Child Protection Plans (up to 28%) following the start of intervention. Looked After Child events also increased to 7%, indicating that although there was a decrease in lower-level safeguarding events (CIN), there was a small group of families who had an increase in higher-level events.

There was an increase in families with school exclusion (up to 12%). There were smaller increases in the percentage of families with criminal offences committed by adults (6%), criminal offences committed by children (8%), LAC events (7%) and NEET members (6%)

8% of families had no events at all in the year following the first intervention date, which indicated a positive change for these families. Where the cluster rules from the 'before' analysis were applied 8% of families would have remained in cluster 10, that is, their circumstances remained the same, and they only had CIN events and school absence in

the year following intervention. For those that changed cluster, 39% would be assigned to cluster 8 (just school absence), which highlights the decrease in CIN events, and 27% to cluster 2 (CPP); the rest were split over the other clusters.

72% of interventions ended in a planned ending, 23% had an unplanned ending and 5% were ongoing a year later. 47% of families received further intervention treatment, which was the second highest percentage over all clusters. Overall, just under a third (31%) of families had an improvement in their circumstances after the start of intervention according to the 'relaxed' criteria; of these, 31% received further intervention treatment. Of the families that showed no improvement (69%), 46% received further intervention treatment.

7.3.6.11 Cluster 11: No Events

Cluster 11 contained 605 families, 471 families (78%) were included in the 'after' analysis

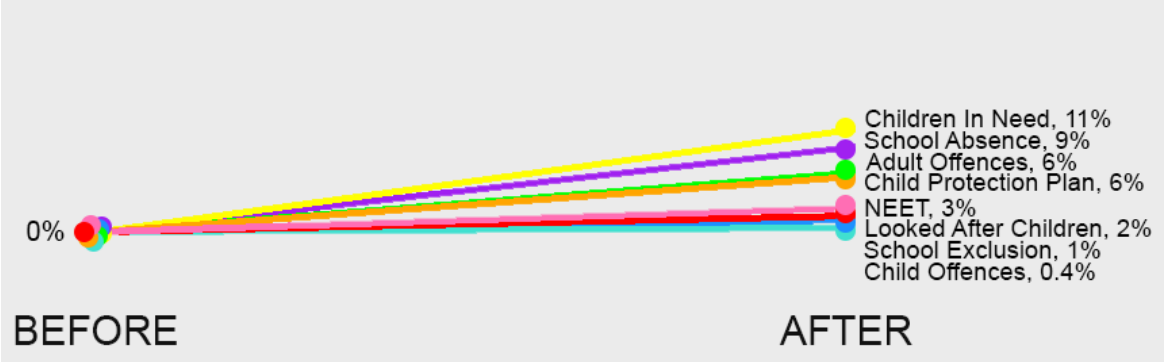


Figure 81: Percentage of families with events in the years before and after the start of intervention for cluster 11

All families in this cluster had no events in the year prior to intervention. However, a small proportion had pre-existing CPPs (3%), LAC events (1%) and members who were NEET (1%). In the year following the start of intervention 73% of families still had no events at all. There were only small increases in the percentage of families with any of the other events, as shown in Figure 81.

In terms of child safeguarding, 11% of families had CIN events, the percentage of families with CPPs increased slightly to 6%, and LAC events increased to 2%. The percentage of families with adult criminal offences was 6%, and 0.4% for child criminal offences. Families with NEET members increased to 3%, and school exclusion was 1%.

73% of families had no events at all in the year following the first intervention date, indicating no change for these families; they had no events in the year before or after the start of intervention. Where the cluster rules from the 'before' analysis were applied, for

those that changed cluster, 6% would be assigned to cluster 8 (just school absence), and 6% to cluster 2 (CPP); the rest were split over the other clusters, except 6 and 7.

75% of interventions ended in a planned ending, 20% had an unplanned ending and 5% were ongoing a year later. 30% of families received further intervention treatment, which was the lowest percentage over all clusters. Overall, almost three quarters (73%) of families had an improvement in their circumstances after the start of intervention according to the 'relaxed' criteria; of these, 28% received further intervention treatment. Of the families that showed no improvement (27%), 47% received further intervention treatment. Cluster 11 consisted of many single person families (just under half were), and almost all of them (94%) had 'improvement'.

7.3.6.12 Summary

Overall, this section has shown that, in the year following the start of intervention, a percentage of families in each of the clusters had some improvement in their circumstances. Where the primary characteristic of each cluster was considered, there was a decrease in the occurrence of those particular events for all clusters (for instance, cluster 2's primary characteristic was that all families had Child Protection Plans; this reduced to 38% of families in the year following the start of intervention treatment). Where the 'relaxed' criteria for 'improvement' (derived from considering the Government guidelines for being 'turned around') were considered, all clusters had families whose lives showed 'improvement' in the year following the start of intervention.

Where the plot for all families (Figure 70) is considered alongside the eleven cluster-level plots, it highlights the extra information gained from analysing the clusters separately. If this plot (and data) alone had been considered, rather than the eleven cluster-level plots, it would have been far less informative, and could not have indicated the underlying complexity of the data. In general, on the 'global' level there appeared to be little change in the percentage of families with events 'before' and 'after', aside from the occurrence of CIN events. Yet, the individual cluster plots highlighted significant changes, with some clusters changing from having all families with particular events 'before', to having very few 'after' (such as cluster 2's change from all families with CPPs 'before' to just over a third of families 'after'). In the overall plot, these significant cluster-level changes were just not visible.

It seems, therefore, that analysing the dataset as a whole had an averaging effect; the large changes that occurred on the cluster-level were simply not represented in Figure 70. This highlights one of the benefits of identifying clusters of similar groups in the data; it allowed a more detailed and insightful analysis of the groups.

7.3.7 Prediction of outcome for families

Machine learning methods were utilised in order to predict the outcome, a year later, for the families. This utilised the data that was known about a family at the start of intervention, that is: the events occurring in the year prior to intervention; any other available data about the family (e.g. family size and type of first intervention treatment received); and the 'place-based' data linked to where a family lived on their first intervention date (as derived in Chapter 6). The models were built in order to identify those factors (or attributes) that had an impact upon the future outcome for a family, and also to determine whether it was possible to predict the outcome, with any level of accuracy, at the start of intervention. For comparison purposes, logistic regression was also utilised in order to determine if it could provide useful information.

Two different sets of models were built: Set 1 considered planned and unplanned endings and whether a family had further treatment as an indication of outcome; Set 2 considered the events occurring before and after the start of intervention for each family and utilised the 'relaxed' criteria of improvement (derived in section 7.3.5) as an alternative indication of outcome.

Decision trees, random forests, generalized boosted models and logistic regression models were built, using the R programming language. The machine learning methods were chosen particularly because they all provide an indication of which attributes are important to a model, and therefore may provide insight into what might be considered a useful predictor of outcome one year after joining the TF programme. They also do not place assumptions upon the data, and many predictor attributes may be utilised if required. Comparing the results of all methods meant that it might be possible to determine if any perform better, or are more suitable, than others.

For each Set, one overall model was built (for all data/families); and separate cluster-level models were also built in order to determine whether model performance might be better on the cluster-level compared to one overall 'global' model. The records in the dataset were split into a training and testing dataset using a 70:30 split; and this split was

also utilised for the cluster-level models. Utilising training and testing datasets provided an extra level of validation as it meant that various settings and different predictors could be experimented with, and the resulting models could be directly compared (as they all utilised the same testing dataset).

Clusters 4 to 7 were excluded from the cluster-level models as it was felt that they contained so few records (each less than 50) that their results may not be useful. This left clusters 1 to 3 and 8 to 11 for analysis. Three sets of predictor attributes were utilised: a large set (Set A) which contained place-based data, event data and anything else that was known about the family at the start of intervention; a smaller set (Set B) which contained fewer attributes; and an even smaller set (set C) with fewer attributes that utilised the cluster assignment (rather than the counts of events). Set C was only utilised for the overall models, not on the cluster level models (as cluster assignment was not relevant here). The largest set (A) was utilised to allow consideration of a wide range of predictors; however, a more parsimonious group of attributes may be easier to understand and so the smaller sets (B and C) were also utilised in order to determine whether they might have similar accuracy. The full list of attributes utilised as predictors is contained in Appendix B, together with model parameters and the full results.

For each model, the baseline accuracy was calculated in order to provide a benchmark; any model that performed with an accuracy on the test dataset that was better than the baseline accuracy might be considered 'good'. Baseline accuracy was calculated by simply considering the accuracy when the most populous category was chosen as the prediction (for example, if 70% of all families had planned endings with further interventions, an accuracy of 70% could be achieved simply by predicting this without even building a model). There were 8 records with missing values (those families that did not link to a postcode); they were excluded from the random forest model and logistic regression models, as the algorithms could not deal with missing values. This meant that the baseline accuracy for these models occasionally differed slightly compared to the other models.

7.3.7.1 Set 1: Predicting planned and Unplanned Endings

This considered planned and unplanned endings, together with whether a family received more than one intervention treatment, as an indication of progress. The previous section (7.3.5) highlighted that the outcome of the first intervention alone did not necessarily

indicate whether a family had shown any improvement with regards to reducing the levels of events that they had, however, it was likely that the outcome of a first intervention may at least be an indicator of cooperation with the TF programme. Also, if a family was referred for more than one treatment type it may imply that they had ongoing or complex issues. The outcome of interventions was considered as there were no other attributes in the data that could indicate participation (or compliance, or progress) with the TF programme; planned or unplanned endings were the only way to confirm that families had actually received some kind of treatment.

The target attribute that was to be predicted therefore had four levels:

- planned ending with no further treatment
- planned ending with further treatment
- unplanned ending with no further treatment
- unplanned ending with further treatment

Where the planned/unplanned ending refers to the first intervention treatment. Since records could only be included that had an outcome for the first intervention (i.e. they had a planned or unplanned ending, and were not still ongoing), this left 2040 records for analysis (or 2032 for the random forest and logistic regression models, excluding records with missing values).

All models were judged by the prediction accuracy on the test dataset. Table 36 details the baseline accuracy compared to the best test set accuracy for each type of model and by cluster assignment. The three different combinations of predictor attributes were utilised (datasets A, B, C) for each model; the model that resulted in the highest accuracy is listed in the table. Highlighted in bold are the models that had a test set accuracy greater than the baseline accuracy, and therefore might be considered 'good'. That is, they picked up some pattern in the data and had accuracy that was better than simply guessing.

Table 36 highlights that none of the models that utilised the full data (i.e. that were not split into clusters) could beat the baseline accuracy. However, there was some success in predicting the outcome for clusters 2, 3, 8, 9 and 10. The boosted models had the most success, followed by the decisions trees and random forests. The logistic regression models did not beat baseline accuracy, although it must be noted that they were misspecified, as many of the predictors were correlated. Overall none of the models had

very high accuracy; most got under half of the predictions correct. And in some cases, the improvement over the baseline accuracy was very small. The models for cluster 10 had the clearest indication of detecting a genuine pattern (both the random forest and boosted model had an improvement over the baseline accuracy of around 10%).

Table 36: Results of models predicting planned/unplanned endings with/without further treatment (with models that beat baseline accuracy highlighted in bold)

	Decision Tree		Random Forest		Boosted model		Logistic Regression	
Cluster	Baseline accuracy	Test accuracy	Baseline accuracy	Test accuracy	Baseline accuracy	Test accuracy	Baseline accuracy	Test accuracy
1	48.8%	41.5%	48.8%	39.0%	48.8%	41.5%	47.7%	44.2%
2	44.3%	38.1%	44.3%	43.3%	44.3%	**48.5%**	42.4%	42.4%
3	41.2%	**44.1%**	41.2%	41.2%	41.2%	**50.0%**	48.5%	48.5%
8	45.2%	-	45.2%	**48.4%**	45.2%	43.6%	41.9%	31.2%
9	45.6%	-	45.6%	44.1%	45.6%	**48.5%**	46.9%	29.7%
10	38.5%	**46.2%**	38.5%	**48.1%**	38.5%	**50.0%**	37.2%	32.6%
11	54.7%	-	54.7%	50.0%	54.7%	54.7%	56.1%	51.1%
All data	47.1%	46.5%	47.1%	43.6%	47.1%	46.8%	47.1%	45.7%

Appendix B contains details of the full results. The following paragraphs summarise the key points for each of the clusters where the models had some success:

Cluster 2: contained 323 families overall, with 97 in the test dataset. Only the generalized boosted model had prediction accuracy (48.5% on the test dataset) that was better than the baseline (by 2.9%). 47 out of 97 records in the test dataset were predicted correctly. The number of families in the test dataset with each outcome, together with the number of correct predictions by the boosted model (in parentheses) was:

Planned ending, no further interventions:	Planned ending, further interventions:	Unplanned ending, no further interventions:	Unplanned ending, further interventions:
43 (31)	37 (15)	13 (1)	4 (0)

Whilst the model appears to have picked up some pattern surrounding planned endings, it struggled to predict unplanned endings, getting only one correct. The most important predictor to the model was the first intervention treatment type a family received, followed by the levels of crime in the area that a family lived. Whilst the model performed a little better than guessing, given it could not accurately predict unplanned endings it may not be very useful.

Cluster 3: contained 113 families overall, with 34 in the test dataset. Both the decision tree (44.1%) and the generalized boosted model (50.0%) had prediction accuracy better than the baseline (41.2%). The boosted model had the best accuracy (just under 9% improvement over baseline) and predicted 17 out of 34 records in the test dataset

correctly. The number of families in the test dataset with each outcome, together with the number of correct predictions by the boosted model (in parentheses) was:

Planned ending, no further interventions:	Planned ending, further interventions:	Unplanned ending, no further interventions:	Unplanned ending, further interventions:
14 (5)	14 (10)	5 (2)	1 (0)

Much like the model for cluster 2, the boosted model appears to have picked up some pattern surrounding planned endings, but struggled to predict unplanned endings. However, it must be noted that there were very few unplanned endings and so these may be more difficult to predict. The most important predictor to the model was the first intervention treatment type a family received, followed by the percentage of single person households in the area a family lived.

The decision tree model, which got 15 out of 34 predictions correct (41.2%), produced a tree with one split: whether a family had the FF intervention treatment type or not. If they did a planned ending with further intervention was predicted, if not a planned ending with no further interventions was predicted. The tree could not predict unplanned endings. The most important predictor to the model was the intervention type.

Both models detected that almost all families who received FF treatment had a planned ending. However, the boosted model had a little more success in terms of further interventions and unplanned endings.

Cluster 8: contained 206 families overall, with 62 in the test dataset. Only the random forest model had prediction accuracy (48.4% on the test dataset) that was better than the baseline (by 3.2%). 30 out of 62 records in the test dataset were predicted correctly. The number of families in the test dataset with each outcome, together with the number of correct predictions by the model (in parentheses) was:

Planned ending, no further interventions:	Planned ending, further interventions:	Unplanned ending, no further interventions:	Unplanned ending, further interventions:
28 (20)	18 (7)	14 (3)	2 (0)

Again, the model appears to have picked up some pattern surrounding planned endings, and did predict a few unplanned endings correctly. The most important predictor to the model was the percentage of people born in Europe living in the family's area, followed by the number of address changes the family had in the year prior to intervention. In contrast to many of the other models, the intervention type had no importance.

Cluster 9: contained 228 families overall, with 68 in the test dataset. Only the generalized boosted model had prediction accuracy (48.5% on the test dataset) that was better than the baseline (by 3.1%). 33 out of 68 records in the test dataset were predicted correctly. The number of families in the test dataset with each outcome, together with the number of correct predictions by the model (in parentheses) was:

Planned ending, no further interventions:	Planned ending, further interventions:	Unplanned ending, no further interventions:	Unplanned ending, further interventions:
31 (26)	23 (7)	9 (0)	5 (0)

The model appears to have picked up some pattern surrounding planned endings, but did not predict any unplanned endings. The important predictors were all place-based, and the most important predictor was the percentage of households that were owned in the area the families lived in, followed by the percentage of people who were Christian in the area, and the percentage who were economically active.

Cluster 10: contained 172 families overall, with 52 in the test dataset. Each of the machine learning models had prediction accuracy that was better than the baseline. The boosted model had the best performance, predicting 26 out of 52 records in the test dataset correctly (50% accuracy, which was an improvement of 11.5% over than the baseline). The number of families in the test dataset with each outcome, together with the number of correct predictions by the boosted model (in parentheses) was:

Planned ending, no further interventions:	Planned ending, further interventions:	Unplanned ending, no further interventions:	Unplanned ending, further interventions:
19 (14)	20 (11)	8 (1)	5 (0)

Again, the model appears to have picked up some pattern surrounding planned endings, but could only predict one unplanned ending correctly. The most important predictor to the model was the first intervention type, followed by the percentage of single person households in the area the family lived, and the percentage of households that were not deprived.

The random forest model had accuracy of 48.1% (almost 10% better than baseline). Again, it mostly predicted planned endings, but got one unplanned ending correct. The most important predictor was the percentage of lone-parent households in the area the family lived, followed by the percentage of people born in the UK. Intervention type had very little importance.

The decision tree model had accuracy of 46.2% (almost 8% better than baseline). It predicted only unplanned endings. The most important predictor was the percentage of households that were owned in the area the family lived, followed by the percentage that were socially rented, and percentage of lone-parent households.

Summary: There was no success in predicting outcome for the global (all data) models, however there was some for the cluster-level models. Whilst some of the machine learning models produced accuracy on the test dataset that was an improvement over the baseline accuracy, most of them had only a small improvement (of just a few percent), and so might not be deemed very useful. However, the models predicting outcome for cluster 10 did perform relatively well. It was clear that most of the models that had some success detected a pattern around planned endings, and were able to distinguish to some degree between those with and without further interventions. However, there was little success in predicting unplanned endings; the random forest and boosted models did manage to predict some without further interventions, but no model could predict those with further interventions. One reason for this may simply be that they were scarce in the dataset.

For some models, the first intervention type that a family received was very important (clusters 2, 3 and 10); aside from this, the over-riding pattern for the models with better performance was that the most important predictors were almost exclusively 'place-based'. That is, they referred to the characteristics of the area that the family lived in, rather than to the family's particular characteristics. This cautiously implies that the events occurring in the year before a family's first intervention (e.g. school absence, CIN events, etc.) did not appear to provide much information as to whether the first intervention would have a planned/unplanned ending with/without further treatment. The higher importance of the 'place-based' attributes suggests that there were at least weak patterns overall regarding where the families lived and the outcome of their interventions.

7.3.7.2 Set 2: Predicting whether a family would have 'improvement'

An alternative method of analysis was to utilise the criteria from section 7.3.5 which considered the outcome for each family in reference to the Government's guidelines on what constituted being 'turned around'. This considered the events occurring in the years before and after the start of intervention treatment. Where a family had a

reduction in the events after the start of intervention, compared to before, it was considered that there had been 'improvement'. The models were therefore built to predict whether or not a family would have 'improvement', and utilised only what was known about the family on the first intervention start date.

Considering 'improvement' was thought a more tangible indication of some sort of progress for the family (compared to the Set 1 predictions), since as noted previously, planned endings did not necessarily correlate with a reduction in the number of events that a family had. The criteria derived for 'improvement' specifically considered the frequency of events. The target attribute utilised the relaxed criteria (defined in section 7.3.5) and had two levels: 'improvement'; or 'no improvement'. There were 1668 records available for analysis (or 1660 for the random forest and logistic regression models, excluding missing values), since only records with one years' worth of available data after the start of intervention could be utilised.

The models were judged by prediction accuracy on the test dataset. Table 37 details the baseline accuracy compared to the test set accuracy for each type of model, and by cluster assignment. Various combinations of predictor attributes were utilised (datasets A, B, C) for each model; the model that resulted in the highest accuracy is listed in the table. Highlighted in bold are the models that had a test set accuracy greater than the baseline accuracy, and therefore might be considered 'good'. That is, they picked up some pattern in the data and had accuracy that was better than simply guessing.

Table 37: Baseline accuracy compared to test set accuracy for models predicting 'improvement'. Models with test set accuracy better than the baseline are highlighted in bold

	Decision Tree		**Random Forest**		**Boosted model**		**Logistic Regression**	
Cluster	Baseline accuracy	**Test accuracy**	Baseline accuracy	**Test accuracy**	Baseline accuracy	**Test accuracy**	Baseline accuracy	**Test accuracy**
1	75.4%	53.6%	75.4%	**76.8%**	75.4%	75.4%	68.7%	68.7%
2	57.3%	-	57.3%	57.3%	57.3%	**64.0%**	53.1%	51.9%
3	71.4%	67.9%	71.4%	67.9%	71.4%	71.4%	81.3%	68.8%
8	65.3%	53.1%	65.3%	61.2%	65.3%	51.0%	51.2%	48.8%
9	54.4%	40.4%	54.4%	50.9%	54.4%	49.1%	63.0%	61.1%
10	68.2%	-	68.2%	65.9%	68.2%	63.6%	68.9%	64.4%
11	72.5%	**73.2%**	72.5%	**74.3%**	72.5%	**76.1%**	68.6%	**70.0%**
All data	54.2%	**67.0%**	54.4%	**62.5%**	54.2%	**65.8%**	54.2%	**64.0%**

Table 37 highlights that there was some success in predicting 'improvement' for clusters 1, 2 and 11. However, the models that had the greatest improvement over baseline accuracy were the ones that utilised the whole dataset (rather than the separate cluster-level models); for all methods, these had test set accuracy that was at least 7% higher

than baseline. Appendix B contains full details of the models that had some success. The following paragraphs summarise the key points for each:

Cluster 1: contained 231 families overall, with 69 in the test dataset. Only the random forest model had prediction accuracy (76.8% on the test dataset) that was better than the baseline (by 1.4%). 53 out of 69 records in the test dataset were predicted correctly. The number of families in the test dataset with each outcome, together with the number of correct predictions by the model (in parentheses) was:

Improvement:	No improvement:
17 (2)	52 (51)

The most important predictor to the model was the number of children in the family. However, it was clear that the model struggled to predict 'improvement' and with such a small improvement over the baseline accuracy, the model performance was only a little better than guessing.

Cluster 2: contained 250 families overall, with 75 in the test dataset. Only the generalized boosted model had prediction accuracy (64% on the test dataset) that was better than the baseline (by 6.7%). 48 out of 75 records in the test dataset were predicted correctly. The number of families in the test dataset with each outcome, together with the number of correct predictions by the model (in parentheses) was:

Improvement:	No improvement:
32 (13)	43 (35)

The most important predictor to the model was the percentage of lone-parent households in the area the family lived in, followed by the percentage of privately rented households in the area, and the number of females in the family. Only these three attributes had any importance. It seems the model did detect some pattern, as it managed to correctly predict a proportion of records with both improvement and no improvement.

Cluster 11: contained 471 families overall, with 142 in the test dataset. All the models had prediction accuracy that was better than the baseline, however only by a small margin (all less than 4%). The boosted model had the best performance, predicting 108 out of 142 records correctly (76.1%). The number of families in the test dataset with each outcome, together with the number of correct predictions by the model (in parentheses) was:

Improvement:	No improvement:
103 (99)	39 (9)

The most important predictor to the model was the number of children in the family, followed by the number of people overall in the family. It seems the model did detect some pattern, as it managed to correctly predict records with both improvement and no improvement. Both the decision tree and random forest models also had number of children and the number of people in the family as the most important predictors. The logistic regression model had no significant predictors ($p < 0.05$).

All data: contained 1668 families overall, with 500 in the test dataset. All the models had prediction accuracy that was better than the baseline. The decision tree model had the best performance, predicting 335 out of 500 records correctly (67%, an improvement of 12.8% over the baseline accuracy). The number of families in the test dataset with each outcome, together with the number of correct predictions by the model (in parentheses) was:

Improvement:	No improvement:
229 (74)	271 (261)

The decision tree had only one split, which was whether a family had children or not. If they did not have children, improvement was predicted; if they did, no improvement was predicted. All the machine learning models identified the number of children in the family, and the number of people overall as the most important predictors. The logistic regression model did not identify these attributes as significant, but did find particular cluster assignments significant. It seems that all the models did detect some pattern, as each managed to correctly predict records with both improvement and no improvement.

Overall, the models picked up that where families had no children, they were more likely to have had 'improvement' (87% of families without children had improvement, compared to 37% of families with children). Equally, where there was only one person in a family, they too were more likely to have 'improvement' (83% of families with one member had 'improvement', compared to 38% with more than one family member).

Summary: Whilst there was some success in predicting 'improvement' for families in clusters 1, 2 and 11, the most notable result was for the models utilising all the data (i.e., all clusters together). All models had test set accuracy that was at least 7% higher than the baseline, indicating that they had detected patterns in the data that were more than

just noise. The most important attributes to the models were the number of children in the family, family size and cluster assignment. Whilst, the 'place-based' attributes still had importance in most of the models overall, they were not considered as important for the Set 2 predictions as they were for Set 1.

7.3.7.3 Predicting whether families with/without children have improvement

Since family size had such importance to the models predicting 'improvement' further experiments were performed where the data was split into those families with children and those without children. It was thought that this might provide further insight into the data. However, whilst a couple of the models had a marginal improvement over the baseline accuracy (around 1%) it would seem that the models were not very useful and could not substantially beat the accuracy attained simply from guessing. The results are contained in Appendix B, part 4.

7.3.7.4 Summary of predictions

The analysis in this section has shown that, overall, whilst it was difficult to predict the future outcome for families, there was some success. Predicting planned/unplanned endings with/without further intervention had some success on the cluster-level, but none on the global level. Whereas, in contrast, predicting 'improvement' had little success on the cluster-level, but some on the global level.

In terms of the Set 1 predictions (planned/unplanned endings), the important factors were the type of intervention treatment a family received, and the data pertaining to where they lived (place-based); the models placed little importance upon the family's characteristics (events, etc.). This partially reflected that different types of intervention treatment had differing success rates, and that the different clusters had varying concentrations of families receiving them.

In terms of the Set 2 predictions, the important factors pertained to family size and the cluster a family belonged in; the place-based attributes had less importance (although still featured heavily). This reflected that, since so much of the data pertains to children, not having children (or having only one member in a family) meant that these families would likely have 'improvement'.

Whilst the importance of the place-based data in many of the models appears to indicate at least weak patterns, this is an area that would warrant further research to more closely consider these patterns.

In terms of the methods employed, the decision tree, random forest and boosted methods each had some success (in terms of test set accuracy); and together helped to confirm the important attributes for some of the models. The logistic regression models had least success, with only two producing accuracy greater than baseline.

7.3.8 Final Summary of clusters

This section (Table 38) draws together the work in the previous chapters and provides a brief summary of the important events for families 'before' and 'after' the start of intervention, for each cluster.

Table 38: Brief comparison of the key characteristics of families in the clusters before and after the start of intervention

Cluster	Year before the start of intervention	Year after the start of intervention
1 (n=291)	High levels of school exclusion and criminal offences (committed by adults and children). Low levels of child safeguarding (CPP, LAC and CIN) events. Lowest level of children attending good/outstanding schools of all the clusters. Families lived in areas with the highest levels (of all clusters) of people born in the UK and of white ethnic group	Decrease in exclusions and criminal offences committed by adults. Decrease in CIN events, however an increase in higher-level child safeguarding issues (CPP and LAC events). 15% of families had no further events after the start of intervention, and overall a quarter of families were considered to have had an 'improvement'
2 (n=335)	Child safeguarding was the main feature. All families had Child Protection events, and there were high levels of CIN and LAC events. Very little school exclusion and criminal offences committed by children. Families tended to have younger children with most aged under 11.	Large decrease in CPPs (down to 38% of families) and CIN events, however a small increase in LAC events. Very little change for the other events. One fifth of families had no further events following the start of intervention, and 43% overall were considered to have had an 'improvement'
3 (n=115)	All families had Looked after Children events. High levels of adult criminal offences compared to other clusters, however low levels of school absence and exclusion. A high proportion of children attended good/outstanding schools compared to other clusters. Families lived in areas with higher levels of economic activity and high population density	Large decrease in LAC events (down to 42% of families) and CIN events, however an increase in CPPs. This suggested serious child safeguarding concerns had decreased for many families. There was an increase in school absence, however the percentage of families with adult criminal offences decreased. 12% of families had no further events after the start of intervention, and 28%

		overall were considered to have had an 'improvement'
4 (n=61)	All families had members who were NEET, and just under a third had criminal offences committed by minors. Low levels of child safeguarding. Compared to other clusters, the families lived in areas that had high levels of household deprivation, social housing and people with no qualifications	Decrease in families with NEET members (down to 63%). Very little change for the other events, although there was an increase in families with CPPs. 10% of families had no further events after the start of intervention, and 35% overall were considered to have had an 'improvement'
5 (n=21)	All families had criminal offences committed by adults, and these were at a high level (with a mean of 4 per family). A third of families had domestic abuse events. There were few child safeguarding (CIN, CPP, LAC) issues	Large decrease in families with adult criminal offences (down to 33%), and a decrease in domestic abuse events. There was an increase in CIN events (which was unusual compared to other clusters), but no high-level safeguarding issues (CPP or LAC). 36% of families had no further events after the start of intervention, and 50% overall were considered to have had an 'improvement' (a higher percentage than all but cluster 11)
6 (n=54)	All families had school absence, at high levels, with families having 39% unauthorised absence on average. There were high levels of school exclusion, and criminal offences committed by children, compared to other clusters. And fewer children attended schools considered as good/outstanding. High levels of child safeguarding issues compared to other clusters	Whilst most (88%) families still had school absence, the average levels were lower (22% unauthorised sessions on average). There was a decrease in the percentage of families with school exclusion, CIN events and criminal offences committed by children. However, an increase in families with NEET members. The prevalence of higher level safeguarding issues remained almost the same. Only 3% of families had no further events after the start of intervention, and 45% overall were considered to have had an 'improvement'
7 (n=25)	All families had criminal offences committed by children, and these were at a high level (mean of 4). Just under half had school absence (but at low levels), and most of those with absence also had CPPs. Families lived in areas with higher levels of social housing compared to other clusters. Proportionately more children attended good/outstanding schools than for any other cluster.	Small decrease in percentage of families with child offences (to 83%). There was an increase in the percentage of families with school absence, exclusion and NEET members. Decrease in CPPs, but increase in LACs, indicating more high-level child safeguarding events. 9% of families had no further events after the start of intervention, and 17% overall were considered to have had an 'improvement'. This was the lowest percentage of all clusters, and reflects that there was an increase in the occurrence of events for many families after intervention

8 (n=223)	All families had school absence but no other events. The average unauthorised school sessions per family was 6%.	Decrease in percentage of families with absence (to 79%), but average levels of absence remained the same (6%). There was an increase in CIN events, and small increases in the other events. 13% of families had no further events after the start of intervention, and 34% overall were considered to have had an 'improvement'
9 (n=243)	All families had CIN events but no other events. The families lived in areas that had the lowest levels of social housing of all clusters, and higher levels of economically active people	Large decrease in the percentage of families with CIN events (to 28%). However, a small increase in CPPs and LAC events, suggesting that although far fewer families had low-level child safeguarding events, a small group had moved to higher-level events. Increase in families with school absence (to 19%) small increases of the other events 45% of families had no further events after the start of intervention, and 45% overall were considered to have had an 'improvement'
10 (n=182)	All families had school absence and at least one CIN event, but no other events. Average percentage of unauthorised absence was 10.6%, a little higher than for most other clusters. The families lived in areas that had the lowest population density, and higher levels of people born in the UK and belonging to the white ethnic group, compared to the other clusters	Large decrease in the percentage of families with CIN events (to 22%), but a much smaller decrease in school absence (to 87%). The average percentage of unauthorised absence per family remained almost the same (10.1%). Increase in families with CPPs, indicating that although there was a decrease in lower-level safeguarding events, a quarter of families (28%) moved to higher-level events. There were small increases in all of the other events. 8% of families had no further events after the start of intervention, and 31% overall were considered to have had an 'improvement'
11 (n=605)	All families had none of the events. 41% of families consisted of single people, a far higher percentage than any other cluster. Half of the families had no children. Families lived in areas with higher levels of household deprivation, and higher levels of people born in the UK and belonging to the white ethnic group, compared to the other clusters	There were small increases in all of the events. 11% of families had CIN events and 9% had school absence. 73% of families had no further events after the start of intervention, meaning that they had no events in the years before and after the start of intervention. 73% overall were considered to have had an 'improvement'

7.4 DISCUSSION

Overall, all of the clusters exhibited change in some aspect in the year following the start of intervention, compared to before. For most clusters, this change was generally positive; each had a group of families who had no further events following the start of intervention. However, for the clusters with no events (cluster 11), one main event (8, 9) and two of the same events (10), there were increases in some events following the start of intervention. This was perhaps inevitable, as in the case of cluster 11, any change could only be negative (there is no improvement upon having no events). However, almost three quarters (73%) of families in cluster 11 had no change; they had no events in the year before or following the start of intervention. Of the families who did have events following the start of intervention, most had only one type of event. Therefore, even where they did develop further issues, they were generally not as complex as those for families in other clusters. As considered in the previous chapter, there may well have been missing data that might explain the families in cluster 11 better, but this analysis could only utilise the available data.

Of all the clusters, the one where many of the families generally appeared to be in a worse situation one year later was cluster 7. The main characteristic of cluster 7 was that all families had criminal offences committed by adults. Whilst there was a reduction in adult criminal offences 'after' (to 83% of families having them), most families had increases in the occurrence of school absence, school exclusion, LAC events and NEET members. However, it should be considered, that cluster 7 was very small, and represented only 1% of families overall.

For the other clusters (1 to 6), all experienced some improvement overall following the start of intervention. The percentage of families with the main characteristics of each cluster (such as CPPs for cluster 2, or LAC events for cluster 3) reduced by a large margin after the start of intervention. Conversation with the ECC suggested that they suspected this effect may have been due to treatment, or keyworkers, focusing closely on the main problem a family had; this was then reduced, but may have in some cases led to other problems receiving less focus.

It is difficult to consider these changes compared to the overall trend of events in the ECC area (Figure 56 to Figure 63) because each family followed a different timeline (that is, they started intervention on different dates). However, over the whole population of the

city from mid-2011 to mid-2015: crimes (committed by adults and children) were generally decreasing; CPP, LAC and NEET events were loosely increasing; school exclusions were decreasing (but increasing from mid-2014); and school absence stayed roughly the same. The CIN events fluctuated, as discussed in section 7.3.2. One could argue that changes that went against these trends (such as the decrease in CPP and LAC events for clusters 2 and 3) might be indicative of the effect of treatment, however, this would require far more detailed analysis of the individual timelines of each family and also a consideration of how previous events might affect future events. In general, close consideration of each family's particular timeline of events may make a useful avenue for future research.

Overall there appears to be no academic literature on applying data mining methods to the TF data; this may be because it is unusual to have access to so much of the underlying data. Much of the existing academic research into the TF programme utilises the overall statistics supplied by the Government (or local councils), or else utilises qualitative data, such as interviews.

However, there are examples of the use of machine learning methods on this data from within some of the Councils involved. At least five local authorities in England have been utilising machine learning methods in order to try and predict children at risk of child abuse (Adams, 2018; McIntyre and Pegg, 2018). A private company called Xantura provides this service for at least two of the Councils; the company also provide a software product, to over 70 Councils across the UK, that utilises data sharing, visualisation and predictive analytics (Xantura, 2018a). As part of this, their software is used to inform the TF programme, although it is not clear how many Councils utilise the software for this purpose. Xantura claim that the software can: identify potential TF and 'maximise claiming' for the Payment by Results scheme; predict how quickly families will be 'turned around' and identify those who will take longest; help with allocation of staff and strategic need priorities; provide greater understanding of the TF generally, and provide insight into where treatment is helping families and where families struggle to meet outcomes (Xantura, 2018b).

McIntyre and Pegg (2018) make the point that at a time when there are large decreases in funding (government funding for local councils will have been cut by £16 billion by 2020), Councils may be adopting predictive systems in order to save money. This may cause

concerns in terms of data privacy and ethical implications over how decisions are made (Pegg and McIntyre, 2018). Xantura provide little detail of their methods, therefore it is not clear what kind of predictive analytics, or data mining, methods that they might be utilising. This lack of transparency underlines the importance of academic work in this area, which may, in contrast, be more transparent and accountable. It may also be subject to more rigorous ethical procedures. This also relates to the wider point made in section 5.2, which was that social scientists might be at risk of being left behind where these large, and socially important, sources of data are concerned; it may be only those with 'data' or machine learning expertise that are asked to analyse them.

Something that was missing from this data (and therefore the analysis) was data around how the families felt, or what they might think of their experience in the TF programme. As considered in section 6.1.3 families may not know they are classed as 'troubled', or that they have been 'turned around'. Data on how the families felt, and what they thought of their time on the TF programme could provide an interesting avenue for further research. It may also be interesting to consider the keyworker opinions. If such data were available, and could be joined to the type of data utilised in this analysis, it could be contrasted (with the counts of events, etc.) to determine whether how families felt had any relation to or effect on their situation. Something that would also be very useful for any future analysis is a definitive attribute detailing where the families had been 'turned around'.

As considered in section 7.3.5.1, the identification of a comparison group in order to consider whether any changes could be directly attributed to the TF programme would be useful. Discussion with ECC found that identification of a comparison group was difficult; the ECC felt that families that were considered in need of help were generally identified and provided with help. Therefore, there was not an obvious large existing group of families in the data with similar needs to the TF who had not received treatment, and so who could be utilised as a comparison group. The previous chapter considered that there was much underlying data missing from this analysis (pertaining to health, anti-social behaviour, receipt of benefits, police call-outs, etc.) which may have helped explain some of the underlying contexts of the families (and clusters); it is possible that were some of this data to become available and aid greater understanding of the families, it might also aid in identifying some kind of comparison group.

The government report into Phase II of the programme also notes difficulties nationally in selecting a comparison group (Department for Communities and Local Government, 2017). It states that as each Local Authority is responsible for selecting their own comparison sample, there can be wide variability in how this is performed (and therefore in how complex the needs of comparison families are) as there is no standardised approach. As of December 2017, work was still ongoing to identify a matched comparison group that was not biased (i.e. that consisted of families not already receiving a similar intervention type of service, and that was random).

7.5 CONCLUSION

This chapter built upon the analysis of the previous chapter, which had explored the Troubled Families data and discovered eleven different clusters of families based upon the events happening in the year prior to first intervention. The clusters had a focus on child safeguarding, education and crime, as this was the data that was available. This chapter considered the events that happened to the TF in the year following the start of their first intervention treatment.

The analysis highlighted that there were changes in the occurrence and types of events for many of the families in the year following their introduction to the TF programme. Analysis on the cluster-level indicated great change within many of the clusters. However, had the analysis been performed only on the 'global' level, it would have seemed that very little had changed, as the overall average of the occurrence of many of the events stayed almost the same 'after' compared to 'before'. This effect underlined the importance of the cluster-level analysis and the extra information that was gained by identifying the clusters of families within the data

Consideration was given to the Government's criteria of what constituted a family to be considered 'turned around'. However, the relevant data was not available to accurately evaluate these criteria, therefore a new set of criteria were derived by considering the data that was available for analysis. These criteria considered a family to have had 'improvement' where they had a decrease in countable events, and had not had any new (different) events occur after the start of intervention. Overall, just under half of families (46%) had 'improvement' in the year following their introduction to the TF programme. The percentages varied widely by cluster, underlining the diversity of the clusters (only

17% from cluster 7 had 'improvement', whereas 73% from cluster 11 did). Overall, the analysis found that having a Planned Ending for a first intervention treatment was not indicative of whether a family had shown any 'improvement', however proportionately more families who did not have 'improvement' were referred for further intervention treatments. It seemed, therefore, that receiving more than one intervention treatment was representative (at least in some cases) of ongoing problems.

Machine learning and regression methods were utilised to determine whether it was possible to predict the outcome for families (that is, whether or not they would have 'improvement' in the year following the start of intervention) given only what was known about them at the start of intervention. If it was possible, it might aid in identifying the important factors, and so provide some understanding of what could lead to families having a good outcome. The models had a little success, and indicated that family size and where a family lived were important. More particularly, that families with children were less likely to have 'improvement'; this is because the focus of the data was on attributes that generally pertained to children (safeguarding and school issues). In terms of accuracy, the machine learning methods consistently outperformed the regression methods; it is likely that this is because they were better suited to the data (which had many predictors, and some correlated attributes).

More broadly the work in this chapter has shown that an averaging effect does exist where analysing the data as a whole. Given that this is just one local authority's data, this suggests there could be wider implications for the analysis of the programme across the whole country. It is likely that this too will be subject to an averaging effect. The Evaluation report (Day et al., 2016) did make the point that there could be an averaging effect; that is, poorer-performing Councils may cancel out well-performing Councils. But this also applies on the family level; families that show no improvement may cancel out those that do show improvement. Identifying clusters, or groups of similar families, across the whole programme may aid in identifying pockets of families where significant positive (or negative) changes have occurred. Identification of these groups might provide deeper insight into their particular problems and aid in understanding what is working within the Programme and what is not.

Overall, the analysis in Chapters 6 and 7 has highlighted that utilising machine learning techniques on this complex, interlinked social data allowed a more detailed analysis and a

far greater understanding than might have been achieved had these methods not been employed. The discovery of the hidden groups within the data meant that each cluster could be explored individually, providing deeper insight into the types of families that existed. The analysis has also highlighted that visualisation techniques frequently utilised in data mining, such as heatmaps, alluvial plots and slopegraphs can further aid in understanding the data and the results of models.

8 CONCLUSION

8.1 INTRODUCTION

This chapter summarises the work contained in this thesis, and considers the research questions that were posed, the contributions to knowledge and ideas for future research.

8.2 SUMMARY OF THE WORK

The main aim of this work was to understand how machine learning techniques might be optimally utilised on the complex, interlinked 'real-life' data that is often used in social science research. The work attempted to understand whether machine learning techniques could effectively facilitate the analysis and comprehension of large social science datasets, and it also aimed to determine which methods were most effective for discovering hidden patterns within these complex, and often noisy, datasets.

In considering this, the more established methods that are arguably most commonly employed in social science research, OLS linear regression and Null Hypothesis Significance Testing, were examined. Linear regression can be an intuitive and powerful tool for social science research in that results are relatively easy to understand and relationships within the data may be identified and quantified. However, linear regression's use in social science research has received sustained criticism over the years, with much of the criticism concerning misuse of the method, and the various misconceptions around its usage (the method itself is not criticised). Many of the concerns centre upon the fact that to be effective, linear regression relies upon strict statistical assumptions. These assumptions can be so difficult to satisfy that they are frequently not adhered to. This may be a particular problem for social science research because the complexity of much social data is difficult to account for; there may be many interactions or hidden groups in the data that are hard to identify. The literature indicated that difficulties satisfying regression assumptions are often not acknowledged. Yet, if a model is misspecified and regression assumptions are not satisfied, the resulting models may not capture the relationships in the data; this can lead to errors and mean that wider inferences made from the models may not be accurate. Overall this may result in inconsistent research. Whilst there are more robust methods that might address some of the problems, the literature indicated that they are frequently not employed.

The research also highlighted that, whilst NHST plays a crucial role in social science research, it is often misunderstood and misapplied, and that placing too much importance upon the results of NHST can distract from providing statistically sound analysis. In particular, the critics of NHST emphasised that the dichotomous nature of the test (accept or reject), based upon an arbitrary significance level could encourage complacency in research. The emphasis upon achieving 'significance' (a statistically significant result is often seen as a pre-requisite for research publication), together with misspecified regression models can lead to research that is unreliable. This might mean that results are not replicated, or that many conflicting results are produced. Overall, this means that research results as a whole are less trusted, and may undermine research quality.

However, this does not mean that the answer to these problems should simply be to stop utilising NHST, or regression methods, rather it indicates that, given some of the weaknesses and misuses of these methods, a wider range of methods should be considered. Greater focus should be placed upon choosing methods that suit the data and the research question. Many data mining methods are non-parametric and allow the use of more diverse data (such as that with different distributions, missing values, mixed data, and 'big' data). In particular, the use of cluster analysis and decision tree analysis was explored; these methods may be utilised to identify hidden groups, interactions and important predictors in data. More generally, machine learning methods also tend to outperform more established methods in terms of predictive ability.

One method that should be adopted from data mining methodology (whether or not the algorithms themselves are adopted) is the idea of testing a predictive model on previously unseen data; this is fundamental to any data mining project. It would be unthinkable in the field of data mining, not to validate a model in this way; it provides confirmation (or not) that any patterns detected are not simply confined to the data sample (or due to random noise) and that they will generalise to the wider population. Yet, as the literature indicated, most social science research does not employ any kind of cross-validation or out of sample testing. In many cases there is no technical reason why it could not be applied (for example, linear regression models could easily be utilised with cross-validation, or a test dataset). The growing availability of 'big' data also means that model validation is increasingly important, as traditional statistical methods and sample sizes are

not necessarily suited to these large datasets. With a large enough sample, there will be many 'significant' results at the $p < 0.05$ level; however, the use of model validation might mean that spurious relationships are easier to identify. In general, if model validation were utilised more frequently it may help to identify any problems early on in an analysis and could provide an extra layer of credibility to research.

Whilst data mining has historically been viewed with suspicion by some within the social sciences, the research on the history and development of data mining highlighted that some of its techniques (such as decision trees) were originally developed by social scientists. They were developed with the aim of providing methods that could enable better understanding of complex data, and that required fewer statistical assumptions, as it was felt that existing regression methods were not suitable for the types of problems and the complex, inter-related data that social scientists frequently dealt with.

One of the reasons that data mining has been viewed with suspicion over the years is the idea of data dredging; that is, searching through the data with no hypothesis and finding misleading results. However, this is not unique to data mining (it is equivalent to p-hacking) and can be a problem with all methods of analysis. Because data mining methods are varied and flexible, they might be utilised to analyse data in order to suggest hypotheses (that is, for exploratory data analysis), and then also to test these hypotheses. As long as research is conducted rigorously (for example, carefully satisfying any parameters and utilising model validation), this need not be a problem.

Another reason that data mining methods have not been widely adopted in social science research is the fact that there are so many algorithms, and no over-riding framework of how to do things. This may mean that it is difficult to know which method is best, and so it may require some experimentation. In any field there is always a learning curve, however, this may be steeper in data mining (simply because of the breadth of methods available). Examining how other (similar) research has been carried out may help with this problem. Other problems relate to improper model validation, poor implementation, or misinterpretation of results, but this may happen in any field; misuse of methods is not unique to data mining. Another concern is that whilst some data mining methods can produce very high predictive accuracy, the results are not always understandable; however, the results of single decision trees and cluster analysis can be easier to interpret and therefore these methods may be of more use to social scientists.

The work in this thesis identified that although there are examples of the use of machine learning techniques within the social sciences, their usage is not widespread and there are large gaps in the literature. However, there is some evidence of the growing use of these methods, for example, in the field of Computational Social Science (CSS). In general, CSS tends to focus on 'big data'; less research is performed on the 'smaller' data, such as social survey data which is synonymous with social science research. Yet there is no technical reason for this, machine learning methods perform well on small datasets. Where machine learning methods have been utilised in social science research, the general consensus was that these methods could provide a useful complement to existing, more established, methods such as regression, and that the non-parametric, flexible nature of the methods meant that problems could be considered from a different perspective.

However, despite the increasing adoption of machine learning methods, this still accounts for only a small portion of social science research overall, and the point was made in the literature that much of this research does not appear in the more traditional social science journals. The point was also made that by not exploring machine learning and more data-driven methods, social scientists are in danger of being left behind, and their research being deemed less relevant. Those who are proficient in machine learning techniques (such as computer or data scientists) may be the ones who are generating social theory and interesting research, because they are taking advantage of new 'big' datasets and utilising the full range of methods that can be applied to them. Yet they may lack the expert knowledge that social scientists have in terms of understanding findings, or in generating relevant research questions. It is vitally important therefore that social scientists are fully involved in this research.

The work carried out in the Case Study chapters aimed to understand whether a data-driven approach, which utilised machine learning methods (such as cluster analysis and decision trees) could effectively analyse and understand a large, complex interlinked social dataset. The dataset was obtained from an English City Council (ECC) and pertained to the Troubled Families Programme in that city. The analysis aimed to determine:

- Whether there existed unique groups (clusters) of families within the Troubled Families data

- Whether the identification of these groups provided deeper insight than one overall analysis of the data might have provided
- How the lives of the families in each cluster changed following their introduction to the TF programme, and whether it was possible to predict, or identify important factors, that may indicate where positive future outcomes would occur

The analysis identified 11 clusters of families, all with different characteristics. The data-driven clustering model utilised only the attributes that were considered complete (that is, did not have known missing data) for the year before the start of a family's first intervention. Therefore, the cluster analysis had a focus on child safeguarding (Children in Need, Child Protection Plans, and Looked After Children), crime (committed by children under 18, or adults), and education (school absence, school exclusion and individuals who were Not in Education, Employment or Training). Some of the clusters were more cohesive than others, but each cluster had their own particular characteristic (such as all families with high levels of school absence, or with CPPs, or no events, etc.).

Geographical analysis of the families in relation to where they lived at the start of intervention indicated that they tended to live in areas with higher percentages of lone-parents, higher levels of deprivation, lower educational levels, poor health, less economic activity and higher levels of social housing. To some degree this might have been expected, but the analysis also indicated subtle patterns on the cluster level. Families whose main issue was child safeguarding tended to live in areas with higher economic activity and higher population density; whereas families with NEET members tended to live in areas with lower economic activity and higher deprivation. Patterns such as this would warrant more detailed further research.

The work showed that decision tree analysis can be effectively utilised to provide understanding of data. Where employed to describe the cluster assignments, the resulting tree visualisation and rules were understandable and also produced a re-usable model that could be used to assign future families to clusters (if necessary). The cluster and tree analysis indicated that child safeguarding issues and school absence were most 'important' to the data; they characterised the clusters. This is likely because they were the most prevalent issues.

The identification and analysis of the clusters showed that, where focussing on child safeguarding, school issues and crime, there were some very cohesive groups in the data.

For instance, those families with just school absence, or no events at all. This indicated that, in terms of just those issues (the eight events considered) many families did not have a wide range of problems. Just over a quarter (28%) had none of the events and 81% had two or fewer different types of events. This lack of diversity of events for many families, and the fact that families must have multiple different events to qualify for the TF programme (according to the Government guidelines) indicated that there were other underlying problems that were simply not represented by the available data. The literature indicated these were likely to be health problems, police call-outs, anti-social behaviour, domestic abuse, and drug/alcohol dependency.

The fact that some of this data was not available meant that it was difficult to analyse the data in relation to the Government guidelines of what constituted a TF, and also in terms of whether a family could be considered 'turned around'. However, the data for the families in the year following their introduction to the TF programme was also analysed. This compared the events a family had in the year prior to joining the programme to the events they had in the year following their introduction to it. Overall, all of the clusters exhibited change in some aspect in the year following the start of intervention. However, had the families been considered as one large group, the averaging effect meant that it was not evident that much change had occurred; on the cluster level it was clear that there were large changes for some families.

Since it was not possible to assess the data in terms of the Government guidelines for being 'turned around', a new criterion of 'improvement' was created utilising the available data. These criteria considered a family to have had 'improvement' where they had a decrease in countable events, and had not had any new (different) events occur after the start of intervention. Overall, just under half of families had 'improvement' in the year following the start of intervention. And 'improvement' varied widely by cluster. In many cases, the event that was the main feature of the cluster had a large reduction one year later (for instance, all families in cluster 2 had child protection plans; one year later just over a third did). However, just over a quarter of families had more, or an increase in events following the start of intervention, and the other quarter had little change.

Overall, the analysis found that having a Planned Ending for a first intervention treatment was not indicative of whether a family had 'improvement'. However, proportionately

more families who did not have 'improvement' were referred for further intervention treatments. It seemed, therefore, that receiving more than one intervention treatment was representative (at least in some cases) of ongoing problems. It is likely that a year may not be a long enough period of time to realistically consider the effect of the programme on the families; many of the problems they face may take a longer period than this to resolve. Further research might consider a longer time-frame.

There was a little success in predicting the outcome ('improvement') for families one year later, given only what was known about them at the start of intervention treatment. The particular machine learning methods were chosen because they can identify attributes that are important to a prediction, and therefore might indicate the factors that are important to families making progress. Various models were built, utilising different methods (decision trees, random forests, boosted models and regression), and on the whole dataset and cluster-level data. In general, the tree-based models performed better (in terms of predictive accuracy) than the regression models; this may be because they were more suited to the data (which contained correlated attributes, and many predictors). For some of the models there were small improvements upon baseline accuracy (that is, they were better than guessing). Where there was success, the models indicated that family size and where a family lived were important. They also indicated that families with children were less likely to have 'improvement'; this is most likely because the focus of the data was on events that generally pertained to children (child safeguarding and school issues). Overall, this part of the analysis highlighted that it is difficult to predict, with a great degree of accuracy, how joining the TF programme might turn out for families, which was not surprising. However, the machine learning methods identified attributes which were important to these predictions; in particular, the fact that the demographic details of where a family lived consistently had importance to the models was interesting and may warrant further research.

Overall, this analysis of the TF data highlighted the difficulties in analysing such complex interlinked data (such as dealing with missing data and hidden underlying contexts), but it also indicated the difficulties in determining the effect of intervention treatment (such as identifying a valid comparison group). However, even with these difficulties the analysis has shed some light on the families and the problems that they face, together with the various issues with the data itself.

8.3 THE RESEARCH QUESTIONS

The following sections consider the research questions that were posed for this thesis:

Can machine learning techniques be effectively utilised or adapted to facilitate the analysis and comprehension of large social science data sets?

The case study work illustrated that machine learning methods can be effectively utilised on large social science datasets. Whilst the TF data itself contained only a few thousand families and so might be considered fairly small, the overall database was large (pertaining to the population of the whole city) and the data was also very wide (in that it contained many attributes). The cluster analysis identified unique groups of families, each with their own particular characteristic. Identifying these groups meant that far greater understanding was provided than would have been were the data considered as a whole. The use of decision trees and visualisation methods meant that the complex data could be presented in a comprehensive and understandable way.

Which data mining methods are most effective for discovering otherwise hidden patterns within complex and often noisy social data?

The analysis showed that cluster analysis and tree-based methods can be effective. The cluster analysis discovered hidden groups of families, and the decision tree analysis provided a set of clear rules to define the clusters. Whilst there were mixed results in terms of utilising the machine learning methods for predictions, this stemmed from the fact that what the models were asked to predict was difficult. However, it was still possible to gain insight into the predictions by considering the important predictors. Tree-based methods are particularly useful because they provide information as to which are the most important factors to a target attribute; this makes them ideally suited for usage in social science research compared to the more 'black-box' machine learning algorithms (such as neural networks). The research highlighted that although 'black-box' or ensemble methods might provide better predictive accuracy, it may be difficult to understand how predictions are made and therefore what is important to the model.

Can data mining methods provide a detailed picture of trends and patterns within a dataset?

The case study analysis included many visualisations in order to indicate the trends and patterns that had occurred within the dataset. Visualisation methods commonly

associated with data mining, such as t-SNE, slopegraphs, heatmaps, alluvial plots and nightingale plots were utilised to illustrate the various patterns around the cluster assignments, and to demonstrate the changes that occurred whilst the families were on the TF programme. The use of visualisations, both on the cluster-level and on the overall data, allowed insight and understanding that may not have been achieved by simply considering summary statistics.

Can machine learning methods be utilised to suggest new hypotheses and research questions?

The analysis of the TF data was far-ranging, considering not only the events that had occurred for a family and what was known about them, but also the characteristics of where they lived. This meant that new research questions were suggested by the data throughout the analysis. Many of these pertained to whether the 'place-based' data (where the families lived) had any effect on the families. Therefore, the most obvious research questions pertain to the place-based data, such as: do families with child-safeguarding events tend to live in areas with higher economic activity and population density? However, other research questions were also suggested by the analysis, such as: where school absence and CIN events are concerned, does having CIN events mean that some families are more likely to have higher levels of school absence? This was suggested from considering clusters 8 (just school absence) and 10 (just school absence and CIN events). Cluster 10 had higher average levels of school absence than cluster 8, and on the surface the only difference between the two was that families in cluster 10 had at least one CIN event (although there may well have been other underlying factors).

8.4 Summary of Contributions

Perhaps the most significant contribution of this research is that it has shown that machine learning techniques can be effectively utilised on a complex, interlinked social data set. It has shown that it is possible to utilise a data-driven approach, with methods such as cluster analysis, decision tree methods and visualisation to gain a better understanding of complex data.

The following summarises the other original contributions of the research:

- Utilising data mining methods on the TF data was unique; there appears to be no academic research that has utilised these methods on the TF data

- The research illustrated that there are large gaps in the existing literature and that machine learning methods are not frequently utilised by social scientists in many subject areas. Yet, this thesis has shown that they can be effectively applied to social data and may also provide a different perspective to a research problem and data
- Provided an in-depth analysis of the available TF data for a particular English City, and identified clusters of families with different characteristics. It illustrated that had all these families simply been analysed as one large group, an averaging effect would have occurred, and many changes and patterns would not have been evident. More broadly, this may have implications for the overall analysis of the TF programme in England; if an averaging effect is present for just one city, this also likely to be present where the data is analysed for the whole country. It may be that identifying clusters of similar families could aid in analysis of the overall TF data, this may uncover patterns that were not evident on the global level
- The research highlighted that visualisation methods more closely associated with data mining, such as t-SNE, alluvial diagrams, heatmaps, slopegraphs and nightingale plots can be effectively utilised to display complex social data
- The research discovered problems with the available ECC data, such as missing data, duplicate people, and likely missing links between people. The ECC found this useful as they had not been aware of some of these problems and it helped them consider how to fix them, what might have been causing them, and any effect it had on their own analysis
- Informed the ECC of the clusters and patterns in their own data, and more generally, the exposure to methods such as clustering means that the ECC now utilise them in their own research

8.5 Directions for Future Work

Considering the case study, there would certainly be much further work in exploring the Troubled Families data. It was clear that many families may have had other problems that were simply not represented by the available data (such as health problems, police call-outs, anti-social behaviour, etc.), were any of this data to become available it could be added to the analysis to consider how it affects the existing clusters, or whether new clusters might exist. In general, the analysis generated questions that could be

investigated were further data made available. These might include: analysis using a control group of families in order to determine whether any improvement could be directly attributed to the TF programme; a closer focus on the 'place-based' element of the data in order to understand how where a family lived may have affected future outcome; the consideration of a longer timeline after the start of intervention treatment; consideration more generally of complete timelines of events for families, to determine whether there were any patterns or particular recurrent sequences of events for them.

More broadly, this thesis has shown that machine learning methods, such as clustering and decision tree learning, can be effectively applied to social data. These methods were particularly chosen because they are interpretable; clusters can be described, and trees can (generally) be plotted. Future work might also include analysis using the methods that are considered 'black box', such as neural networks. That is, methods that are considered less interpretable; analysis would consider whether such methods can also provide some kind of useful explanatory power in social science research.

8.6 Final Thoughts

In conclusion, this thesis has shown that machine learning methods can be effectively utilised on social science datasets and that these methods can be used for far more than just prediction. They offer ways to identify important predictors in a dataset, provide a better understanding of the structure of the data, identify hidden groups and relationships and aid in generating research questions and hypotheses. However, this need not mean that machine learning methods should replace the more established methods, rather their use alongside these methods may provide enhanced understanding, model validation, and deeper perspective. Perhaps in future years, machine learning methods will be considered a standard part of the wider toolkit of social science methods.

REFERENCES

[Author withheld] (2017) *Troubled Families - Interventions*. [ECC website]. [Online] [Accessed on 29th January 2015] [ECC website].

Abbott, A. (2007) 'Notes on Replication.' *Sociological Methods & Research*, 36(2) pp. 210–219.

Abelson, R. P. (1997a) 'A Retrospective on the Significance Test Ban of 1999 (If there were no significance tests, they would be invented).' *In* Harlow, L. L., Mulaik, S. A., and Steiger, J. H. (eds) *What if There Were no Significance Tests?*, pp. 117–141.

Abelson, R. P. (1997b) 'On the Surprising Longevity of Flogged Horses: Why There Is a Case for the Significance Test.' *Psychological Science*, 8(1) pp. 12–15.

Acharya, A. and Sinha, D. (2014) 'Early Prediction of Students Performance using Machine Learning Techniques.' *International Journal of Computer Applications*, 107(1) pp. 37–43.

Achen, C. H. (1977) 'Measuring Representation: Perils of the Correlation Coefficient.' *American Journal of Political Science*, 21(4) pp. 805–815.

Achen, C. H. (1990) 'What Does "Explained Variance" Explain?: Reply.' *Political Analysis*, 2(1) pp. 173–184.

Achen, C. H. (2005) 'Let's put garbage-can regressions and garbage-can probits where they belong.' *Conflict Management and Peace Science*, 22(4) pp. 327–339.

Acxiom (2017) *Personicx: About*. [Online] [Accessed on 28th February 2017] http://www.personicx.co.uk/about.html.

Acxiom Corporation (2014) *Acxiom Corporation Annual Report 2014*.

Adams, J. (2018) *Councils using 'hundreds of thousands of people's data to try and predict child abuse.'* The Telegraph. [Online] [Accessed on 18th September 2018] https://www.telegraph.co.uk/news/2018/09/16/councils-using-hundreds-thousands-peoples-data-try-predict-child/.

Agarwal, S., Pandey, G. N. and Tiwari, M. D. (2012) 'Data Mining in Education: Data Classification and Decision Tree Approach.' *International Journal of e-Education, e-Management and e-Learning*, 2(2) pp. 140–145.

Alberts, B., Cicerone, R. J., Fienberg, S. E., Kamb, A., McNutt, M., Nerem, R. M., Schekman, R., Shiffrin, R., Stodden, V., Suresh, S., Zuber, M. T., Pope, B. K. and Jamieson, K. H. (2015) 'Self-correction in science at work.' *Science*, 348(6242) pp. 1420–1422.

Anderson, C. (2008) *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. Wired. [Online] [Accessed on 3rd March 2017] http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

Ang, R. P. and Goh, D. H. (2013) 'Predicting juvenile offending: a comparison of data mining methods.' *International journal of offender therapy and comparative criminology*, 57(2) pp. 191–207.

Anscombe, F. J. (1973) 'Graphs in Statistical Analysis.' *The American Statistician*, 27(1) pp. 17–21.

Armstrong, J. S. (2007) 'Significance tests harm progress in forecasting.' *International Journal of Forecasting*, 23(2) pp. 321–327.

Armstrong, J. S. (2012) 'Illusions in regression analysis.' *International Journal of Forecasting*, 28(3) pp. 689–694.

Assi, J., Lucchini, M. and Spagnolo, A. (2012) 'Mapping patterns of well-being and quality of life in extended Europe.' *International Review of Economics*, 59(4) pp. 409–430.

Azevedo, A. and Santos, M. F. (2008) 'KDD, SEMMA and CRISP-DM: a parallel overview.' *In IADIS European Conference Data Mining*, pp. 182–185.

Bai, L., Liang, J., Dang, C. and Cao, F. (2011) 'A novel attribute weighting algorithm for clustering high-dimensional categorical data.' *Pattern Recognition*. Elsevier, 44(12) pp. 2843–2861.

Bakan, D. (1966) 'The test of significance in psychological research.' *Psychological Bulletin*, 66(6) pp. 423–437.

Bambauer, D. E. (2013) 'Privacy Versus Security.' *The Journal of Criminal Law & Criminology*, 103(3).

Barbaro, M. and Zeller Jr, T. (2006) 'A Face Is Exposed for AOL Searcher No. 4417749.' *The New York Times*. [Online] 9th August. http://www.nytimes.com/2006/08/09/technology/09aol.html.

Bate, A. (2016) *The Troubled Families programme (England)*.

Bawden, A. (2015) *Is the success of the government's troubled families scheme too good to be true?* The Guardian. [Online] [Accessed on 11th August 2016] https://www.theguardian.com/society/2015/nov/11/troubled-family-programme-government-success-council-figures.

Bearman, P. (2015) 'Big Data and historical social science.' *Big Data & Society*, (December) pp. 1–5.

Belson, W. A. W. (1959) 'Matching and prediction on the principle of biological classification.' *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 8(2) pp. 65–75.

Benjamini, Y. (2010) 'Simultaneous and selective inference: Current successes and future challenges.' *Biometrical Journal*, 52(6) pp. 708–721.

Berk, R. A. (2004) *Regression analysis: A Constructive Critique*. Thousand Oaks: SAGE Publications, Inc.

Berk, R. A. and Bleich, J. (2013) 'Statistical Procedures for Forecasting Criminal Behavior.' *Criminology & Public Policy*, 12(3) pp. 513–544.

Berk, R. A., Sorenson, S. B. and Barnes, G. (2016) 'Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions.' *Journal of Empirical Legal Studies*, 13(1) pp. 94–115.

Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K. and Zhao, L. (2014) 'Misspecified Mean Function Regression: Making Good Use of Regression Models That Are Wrong.' *Sociological Methods and Research*, 43(3) pp. 422–451.

Berk, R., Brown, L., Buja, A., George, E. and Zhao, L. (2017) 'Working with Misspecified Regression Models.' *Journal of Quantitative Criminology*. Springer US, April.

Berkson, J. (1938) 'Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test.' *Journal of the American Statistical Association*, 33(203) pp. 526–536.

Boddy, J., Statham, J., Warwick, I., Hollingworth, K. and Spencer, G. (2016) 'What kind of trouble? Meeting the health needs of "Troubled Families" through intensive family support.' *Social Policy and Society*, 15(2) pp. 275–288.

Bohanec, M. and Bratko, I. (1994) 'Trading Accuracy for Simplicity in Decision Trees.' *Machine Learning*, 15(3) pp. 223–250.

Borge-Holthoefer, J., Moreno, Y. and Yasseri, T. (2016) 'Editorial: At the Crossroads: Lessons and Challenges in Computational Social Science.' *Frontiers in Physics*, 4(August) p. 37.

Boslaugh, S. (2013) *Statistics in a Nutshell*. 2nd ed., O'Reilly.

Branch, M. (2014) 'Malignant side effects of null-hypothesis significance testing.' *Theory & Psychology*, 24(2) pp. 256–277.

Breiman, L. (1996) ‘Bagging predictors.’ *Machine Learning*, 24 pp. 123–140.

Breiman, L. (2001a) ‘Random forests.’ *Machine learning*, 45 pp. 5–32.

Breiman, L. (2001b) ‘Statistical Modeling: The Two Cultures.’ *Statistical Science*, 16(3) pp. 199–215.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984) *Classification and Regression Trees*. Belmont, California: Wadsworth International Group.

Bruns, S. B. and Ioannidis, J. P. A. (2016) ‘P-curve and p-hacking in observational research.’ *PLoS ONE*, 11(2) pp. 1–13.

Burrell, J. (2016) ‘How the machine “thinks”: Understanding opacity in machine learning algorithms.’ *Big Data & Society*, 3(1) p. 205395171562251.

Burrows, R. J. and Savage, M. (2014) ‘After the crisis? Big Data and the methodological challenges of empirical sociology.’ *Big Data and Society*, (June) pp. 1–6.

Byrne, D. (2009) ‘Using cluster analysis, qualitative comparative Using Cluster Analysis, Qualitative Comparative Analysis and NVivo in Relation to the Establishment of Causal Configurations with Pre-existing Large-N Datasets: Machining Hermeneutics.’ *In* Byrne, D. and Ragin, C. C. (eds) *The SAGE Handbook of Case-Based Methods*. Sage, pp. 260–268.

Calinski, T. and Harabasz, J. (1974) ‘A Dendrite Method for Cluster Analysis.’ *Communications in Statistics - Simulation and Computation*, 3(1) pp. 1–27.

Cameron, D. (2011) *Troubled Families Speech*. [Online] [Accessed on 28th August 2016] https://www.gov.uk/government/speeches/troubled-families-speech.

Carver, R. P. (1978) ‘The Case Against Statistical Significance Testing.’ *Harvard Educational Review*, 48(3) pp. 378–399.

Castellani, B. and Hafferty, F. (2009) *Sociology and complexity science: a new field of inquiry*. Berlin: Springer.

Castellani, B. and Rajaram, R. (2012) ‘Case-based modeling and the SACS Toolkit: a mathematical outline.’ *Computational and Mathematical Organization Theory*, 18(2) pp. 153–174.

Cave, J. (2016) ‘The ethics of data and of data science: an economist’s perspective.’ *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083).

Chang, C.-J. and Shyue, S.-W. (2009) ‘A study on the application of data mining to disadvantaged social classes in Taiwan’s population census.’ *Expert Systems with Applications*. Elsevier Ltd, 36(1) pp. 510–518.

Chang, R. M., Kauffman, R. J. and Kwon, Y. (2014) ‘Understanding the paradigm shift to computational social science in the presence of big data.’ *Decision Support Systems*. Elsevier B.V., 63 pp. 67–80.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000) *CRISP-DM 1.0 Step-by-step data mining guide*. *SPSS inc.*

Chen, H. Y., Hou, T.-W. and Chuang, C.-H. (2010) ‘Applying data mining to explore the risk factors of parenting stress.’ *Expert Systems with Applications*. Elsevier Ltd, 37(1) pp. 598–601.

Chertov, O. and Aleksandrova, M. (2013) ‘Fuzzy clustering with prototype extraction for census data analysis.’ *Studies in Fuzziness and Soft Computing*, 291 pp. 289–313.

Chow, S. L. (1998) ‘Precis of Statistical Significance: Rationale, validity, and utility.’ *Behavioral and Brain Sciences*, 21 pp. 169–239.

Cioffi-Revilla, C. (2010) ‘Computational social science.’ *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3) pp. 259–271.

Coenen, F. (2011) 'Data mining: past, present and future.' *The Knowledge Engineering Review*, 26(01) pp. 25–29.

Cohen, J. (1994) 'The Earth Is Round (P < .05).' *The American Psychologist*, 49(12) pp. 997–1003.

Communities and Local Government (2012) *The Troubled Families programme: Financial Framework for the Troubled Families programme's payment-by-results scheme for local authorities*.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.-P., Sanchez, A., Nowak, A., Flache, A., San Miguel, M. and Helbing, D. (2012) 'Manifesto of computational social science.' *The European Physical Journal Special Topics*, 214(1) pp. 325–346.

Cook, C. (2016) *Troubled Families report 'suppressed.'* BBC Newsnight. [Online] [Accessed on 9th August 2016] http://www.bbc.co.uk/news/uk-politics-37010486.

Cortina, J. M. and Dunlap, W. P. (1997) 'On the Logic and Purpose of Significance Testing.' *Psychological Methods*, 2(2) pp. 161–172.

Couzin-Frankel, J. (2013) 'The Power of Negative Thinking.' *Science*, 342 pp. 68–69.

Cowls, J. and Schroeder, R. (2015) 'Causation, Correlation, and Big Data in Social Science Research.' *Policy & Internet*, 7(4) pp. 447–472.

Crosas, M., King, G., Honaker, J. and Sweeney, L. (2015) 'Automating Open Science for Big Data.' *ANNALS of the American Academy of Political and Social Science*, 659(May) pp. 260–273.

Crossley, S. (2015) *The Troubled Families Programme: the perfect social policy? Centre for Crime and Justice Studies - Briefing Paper*.

Crossley, S. (2018) *Troublemakers: The construction of 'troubled families' as a social problem*. Policy Press.

Cumming, G. (2014) 'The New Statistics: Why and How.' *Psychological Science*, 25(7) pp. 7–29.

Database Marketing - Businessweek (1994) Businessweek. [Online] [Accessed on 29th November 2015] http://www.bloomberg.com/bw/stories/1994-09-04/database-marketing.

Dawson, J. F. (2014) 'Moderation in Management Research: What, Why, When, and How.' *Journal of Business and Psychology*, 29(1) pp. 1–19.

Day, L., Bryson, C., White, C., Purdon, S., Bewley, H., Sala, L. K. and Portes, J. (2016) *National Evaluation of the Troubled Families Programme Final Synthesis Report*.

Delavari, N., Phon-Amnuaisuk, S. and Beikzadeh, M. (2008) 'Data Mining Application in Higher Learning Institutions.' *Informatics in Education*, 7(1) pp. 31–54.

Department for Communities and Local Government (2012) *Working with troubled families. A guide to the evidence and good practice*.

Department for Communities and Local Government (2015) *Financial Framework for the Expanded Troubled Families Programme*.

Department for Communities and Local Government (2017) *National evaluation of the Troubled Families Programme 2015 - 2020: Family Outcomes - national and local datasets*.

Department for Education (2016) *School Performance Tables*. GOV.UK. [Online] [Accessed on 16th December 2016] https//www.gov.uk/government/collections/school-performance-tables-about-the-data.

Dimitriadou, E. and Dolnicar, S. (2002) 'An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets.' *Psychometrika*, 67(1) pp. 137–160.

Domingos, P. (2012) 'A few useful things to know about machine learning.' *Communications of the*

ACM, 55(10) pp. 78–87.

Donohue, J. J. and Levitt, S. D. (2001) 'The Impact of Legalized Abortion on Crime.' *The Quarterly Journal of Economics*, CXVI(2).

Draper, N. R. and Smith, H. (2014) *Applied Regression Analysis*. 3rd ed., John Wiley & Sons, Incorporated.

Duda, R. O., Hart, P. E. and Stork, D. G. (1973) *Pattern Classification and Scene Analysis*. 2nd ed., John Wiley & Sons, Inc.

Duncan, D. F., Kum, H.-C., Weigensberg, E. C., Flair, K. A. and Stewart, C. J. (2008) 'Informing child welfare policy and practice: Using knowledge discovery and data mining technology via a dynamic Web site.' *Child maltreatment*, 13(4) pp. 383–91.

Džeroski, S. (2007) 'Towards a General Framework for Data Mining.' *In* Dzeroski, S. and Struyf, J. (eds) *Knowledge Discovery in Inductive Databases: 5th International Workshop, KDID 2006*. Berlin: Springer Berlin Heidelberg, pp. 259–300.

Efron, B. (1983) 'Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation.' *Journal of the American Statistical Association*, 78(382) pp. 316–331.

Eisenhauer, J. G. (2015) 'Statistical Gestalt: Illustrating Interaction with Indicator Variables.' *General Linear Model Journal*, 41(1) pp. 36–45.

Elwert, F. and Winship, C. (2010) 'Effect Heterogeneity and Bias in Main-Effects-Only Regression Models.' *In Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pp. 327–336.

Engman, A. (2013) 'Is there life after P<0.05? Statistical significance and quantitative sociology.' *Quality and Quantity*, 47(1) pp. 257–270.

Er, E. (2012) 'Identifying At-Risk Students Using Machine Learning Techniques.' *International Journal of Machine Learning and Computing*, 2(4) p. 480.

Erceg-Hurn, D. M. and Mirosevich, V. M. (2008) 'Modern Robust Statistical Methods: An Easy Way to Maximize the Accuracy and Power of Your Research.' *American Psychologist*, 63(7) pp. 591–601.

Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis*. 5th ed., Wiley.

Falk, R. (1998) 'Replication - A Step in the Right Direction.' *Theory & Psychology*, 8(3) pp. 313–321.

Falk, R. and Greenbaum, C. W. (1995) 'Significance Tests Die Hard: The Amazing Persistence of a Probabilistic Misconception.' *Theory & Psychology*, 5(1) pp. 75–98.

Fanelli, D. (2012) 'Negative results are disappearing from most disciplines and countries.' *Scientometrics*, 90(3) pp. 891–904.

Fayyad, U. M., Piatetsky-Shapiro, G. and Ramasamy, U. (2003) 'Data Mining: the next 10 years.' *ACM SIGKDD Explorations Newsletter*, 5(2) pp. 191–196.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From Data Mining to Knowledge Discovery in Databases.' *AI Magazine*, 17(3) pp. 37–54.

Federal Trade Commission (2014) *Data brokers: A call for transparency and accountability*. *Data Brokers and the Need for Transparency and Accountability*.

Finch, H. (2005) 'Comparison of Distance Measures in Cluster Analysis with Dichotomous Data.' *Journal of Data Science*, 3 pp. 85–100.

Fomel, S. and Claerbout, J. F. (2009) 'Guest Editors' Introduction: Reproducible Research.' *Computing in Science & Engineering*, 11(1) pp. 5–7.

Foote, C. L. and Goetz, C. F. (2008) 'the Impact of Legalized Abortion on Crime: Comment.' *The Quarterly Journal of Economics*, 123(1) pp. 407–423.

Franco, A., Malhotra, N. and Simonovits, G. (2014) 'Publication bias in the social sciences: Unlocking the file drawer.' *Science*, 345(6203) pp. 1502–1505.

Freedman, D. (1995) 'Some Issues in the Foundation of Statistics.' *Foundations of Science*, 1(1) pp. 19–39.

Freedman, D. A. (2009) *Statistical Models: Theory and Practice*. Cambridge University Press.

Freese, J. (2007) 'Replication Standards for Quantitative Social Science: Why Not Sociology?' *Sociological Methods & Research*, 36(2) pp. 153–172.

Freund, Y. and Schapire, R. E. (1997) 'A desicion-theoretic generalization of on-line learning and an application to boosting.' *Journal of Computer and System Sciences*, 55 pp. 119–139.

Friedman, J. H. (2001) 'Greedy Function Approximation: A Gradient Boosting Machine.' *The Annals of Statistics*, 29(5) pp. 1189–1232.

Full Fact (2012) *Are 120,000 'problem' families costing the taxpayer £9bn?* Full Fact. [Online] [Accessed on 30th August 2016] https://fullfact.org/news/are-120000-problem-families-costing-taxpayer-9bn/.

Galton, F. (1886) 'Regression Towards Mediocrity in Heriditary Stature.' *Journal of the Anthropological Institute of Great Britain and Ireland*, 15 pp. 246–263.

Gan, G. and Wu, J. (2004) 'Subspace clustering for high dimensional categorical data.' *ACM SIGKDD Explorations Newsletter*, 6(2) p. 87.

Gan, G., Wu, J. and Yang, Z. (2006) 'PARTCAT: A Subspace Clustering Algorithm for High Dimensional Categorical Data.' *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. Ieee pp. 4406–4412.

Gelman, A. and Loken, E. (2014) 'The statistical Crisis in science.' *American Scientist*, 102(6) pp. 460–465.

Gelman, A. and Stern, H. (2006) 'The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant.' *The American Statistician*, 60(4) pp. 328–331.

Gibson, D., Kleinberg, J. and Raghavan, P. (1998) 'Clustering categorical data: An approach based on dynamical systems.' *The VLDB Journal—The International Journal on Very Large Data Bases*, 8(3–4) pp. 1–22.

Gigerenzer, G. (2004) 'Mindless statistics.' *Journal of Socio-Economics*, 33(5) pp. 587–606.

Giles, J. (2012) 'Making the Links.' *Nature*, 488(23 August) pp. 448–450.

Gillo, M. W. and Shelly, M. W. (1974) 'Predictive Modeling of Multivariable and Multivariate Data.' *Journal of the American Statistical Association*, 69(347) pp. 646–653.

Gray, E., Jennings, W., Farrall, S. and Hay, C. (2015) 'Small Big Data: Using multiple data-sets to explore unfolding social and economic change.' *Big Data & Society*, January-Ju(June) pp. 1–6.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. and Altman, D. G. (2016) 'Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.' *European Journal of Epidemiology*. Springer Netherlands, 31(4) pp. 337–350.

Grimmer, J. (2015) 'We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together.' *PS: Political Science & Politics*, 48(01) pp. 80–83.

Gross, J. H. (2015) 'Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance.' *American Journal of Political Science*, 59(3) pp. 775–788.

Grossi, E., Blessi, G. T., Sacco, P. L. and Buscema, M. (2012) 'The Interaction Between Culture, Health and Psychological Well-Being: Data Mining from the Italian Culture and Well-Being Project.' *Journal of Happiness Studies*, 13(1) pp. 129–148.

Guha, S., Rastogi, R. and Shim, K. (2000) 'ROCK: A robust clustering algorithm for categorical attributes.' *Information Systems*, 25(5) pp. 345–366.

Gutierrez, J. and Leroy, G. (2009) 'Using Decision Trees to Predict Crime Reporting.' *In* Siau, K. and Erickson, J. (eds) *Advanced Principles for Improving Database Design, Systems Modeling, and Software Development*. IGI Global, pp. 132–145.

Guyon, I. and Elisseeff, A. (2003) 'An Introduction to Variable and Feature Selection.' *Journal of Machine Learning Research*, 3 pp. 1157–1182.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. and Erlich, Y. (2013) 'Identifying Personal Genomes by Surname Inference.' *Science*, 339(January) pp. 321–325.

Hagen, R. L. (1997) 'In Praise of the Null Hypothesis Statistical Test.' *American Psychologist*, 52(1) pp. 15–24.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*. 2nd ed., Springer.

Hayden, C. and Jenkins, C. (2014) '"Troubled Families" Programme in England: "wicked problems" and policy-based evidence.' *Policy Studies*. Taylor & Francis, 35(6) pp. 631–649.

Hayes, A. F. and Cai, L. (2007) 'Using heteroskedasticity-consistent standard error estimators in OLS regression: an introduction and software implementation.' *Behavior research methods*, 39(4) pp. 709–722.

Healy, K. and Moody, J. (2014) 'Data Visualization in Sociology.' *Annual Review of Sociology*, 40(1) pp. 105–128.

Heiberger, R. H. and Riebling, J. R. (2016) 'Installing computational social science: Facing the challenges of new information and communication technologies in social science.' *Methodological Innovations*, 9.

Hill, D. W. J. and Jones, Z. M. (2014) 'An Empirical Evaluation of Explanations for State Repression.' *The American Political Science Review*, 108(3) pp. 661–687.

Hindman, M. (2015) 'Building Better Models: Prediction, Replication and Machine Learning in the Social Sciences.' *The ANNALS of the American Academy of Political and Social Science*, 659(1) pp. 48–62.

Hofman, J. M., Sharma, A. and Watts, D. J. (2017) 'Prediction and Explanation in Social Systems.' *Science*, 335(6324) pp. 486–488.

Hofmann, S. G. (2002) 'Fisher's Fallacy and NHST's Flawed Logic.' *American Psychologist*, 57(1) pp. 69–70.

Home Office (2016) *ASB Incidents, Crime and Outcomes*. Data.Police.UK. [Online] [Accessed on 3rd October 2016] http://data.police.uk.

House of Commons Committee of Public Accounts (2016) *Troubled families: progress review. Thirty-third Report of Session 2016-17*. London.

Huang, Z. (1997) 'A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining.' *Research Issues on Data Mining and Knowledge Discovery* pp. 1–8.

Hubert, L. and Arabie, P. (1985) 'Comparing Partitions.' *Journal of Classification*, 2(193).

Ioannidis, J. P. A. (2005) 'Why Most Published Research Findings Are False.' *PLoS Medicine*, 2(8) pp. 0696–0701.

J. C. Gower (1971) 'A general coefficient of similarity and some of its properties.' *Biometrics*, 27(4) pp. 857–871.

Jagadish, H. V., Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J., Labrinidis, A., Madden, S., Papakonstantinou, Y., Patel, J. M.,

Ramakrishnan, R., Ross, K., Shahabi, C., Suciu, D., Vaithyanathan, S. and Widom, J. (2012) *Challenges and opportunities with big data*. *White Paper for Computing Community Consortium*.

Jain, A. K. (2010) 'Data clustering: 50 years beyond K-means.' *Pattern Recognition Letters*. Elsevier B.V., 31(8) pp. 651–666.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An Introduction to Statistical Learning with Applications in R*. Springer.

Jiang, S., Ferreira, J. and González, M. C. (2012) 'Clustering daily patterns of human activities in the city.' *Data Mining and Knowledge Discovery*, 25(3) pp. 478–510.

Johnson, D. H. (1999) 'The insignificance of statistical significance testing.' *The Journal of Wildlife Management*, 63 pp. 763–772.

Johnson, J. A. (2012) 'Ethics of Data Mining and Predictive Analytics in Higher Education.' *In Rocky Mountain Association for Institutional Research Conference*. Laramie, Wyoming, pp. 1–24.

Joyce, T. (2004) 'Did Legalized Abortion Lower Crime?' *The Journal of Human Resources*, 39(1) pp. 1–28.

Kass, G. V. (1980) 'An Exploratory Technique for Investigating Large Quantities of Categorical Data.' *Applied Statistics*, 29(2) pp. 119–127.

Keely, L. C. and Tan, C. M. (2008) 'Understanding preferences for income redistribution.' *Journal of Public Economics*, 92(5–6) pp. 944–961.

King, G. (1986) 'How not to lie with statistics: Avoiding common mistakes in quantitative political science.' *American Journal of Political Science*, 30(3) pp. 666–687.

King, G. (1991) '"Truth" Is Stranger than Prediction, More Questionable than Causal Inference.' *American Journal of Political Science*, 35(4) pp. 1047–1053.

King, G. (2016) 'Preface: Big Data Is Not About The Data!' *In* Alvarez, R. M. (ed.) *Computational Social Science: Discovery and Prediction*. Cambridge University Press.

King, M. W. and Resick, P. A. (2014) 'Data mining in psychological treatment research: a primer on classification and regression trees.' *Journal of consulting and clinical psychology*, 82(5) pp. 895–905.

Kirk, R. E. (2003) 'The Importance of Effect Magnitude.' *In* Davis, S. F. (ed.) *Handbook of Research Methods in Experimental Psychology*. Blackwell Publishing Ltd., pp. 83–105.

Kitchin, R. (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE Publications.

Kitchin, R. and McArdle, G. (2016) 'What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets.' *Big Data & Society*, January-Ju.

Kosinski, M., Stillwell, D. and Graepel, T. (2013) 'Private traits and attributes are predictable from digital records of human behavior.' *Proceedings of the National Academy of Sciences of the United States of America*, 110(15) pp. 5802–5.

Kotsiantis, S. B. (2013) 'Decision trees: a recent overview.' *Artificial Intelligence Review*, 39(4) pp. 261–283.

Kotsiantis, S., Pierrakeas, C. and Pintelas, P. (2004) 'Predicting Students' Performance in Distance Learning Using Machine Learning Techniques.' *Applied Artificial Intelligence*, 18(5) pp. 411–426.

Kriegel, H.-P., Kröger, P. and Zimek, A. (2009) 'Clustering high-dimensional data.' *ACM Transactions on Knowledge Discovery from Data*, 3(1) pp. 1–58.

Krijthe, J. and van der Maaten, L. (2017) *Package 'Rtsne.'*

Kuhn, M. (2013) *Predictive Modeling with R and the caret Packages*.

Kuhn, M. and Johnson, K. (2013) *Applied Predictive Modeling*. *Springer*.

Kuroki, Y. (2015) 'Risk factors for suicidal behaviors among Filipino Americans: a data mining approach.' *American journal of orthopsychiatry*, 85(1) pp. 34–42.

Lalayants, M., Epstein, I. and Adamy, D. (2011) 'Multidisciplinary consultation in child protection: A clinical data-mining evaluation.' *International Journal of Social Welfare*, 20(2) pp. 156–166.

Lambdin, C. (2012) 'Significance tests as sorcery: Science is empirical--significance tests are not.' *Theory & Psychology*, 22(1) pp. 67–90.

Lang, J. M., Rothman, K. J. and Cann, C. I. (1998) 'That Confounded P-Value.' *Epidemiology*, 9(1) pp. 7–8.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Alstyne, M. Van (2009) 'Computational social science.' *Science*, 323(5915) pp. 721–723.

Leahey, E. (2005) 'Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology Author.' *Social Forces*, 84(1) pp. 1–24.

Leamer, E. E. (1983) 'Let's Take the Con Out of Econometrics.' *The American Economic Review*, 73(1) pp. 31–43.

Lecoutre, B., Lecoutre, M.-P. and Poitevineau, J. (2001) 'Uses, Abuses and Misuses of Significance Tests in the Scientific Community: Won't the Bayesian Choice Be Unavoidable?' *International Statistical Review / Revue Internationale de Statistique*, 69(3) pp. 399–417.

Lehrer, D., Leschke, J., Lhachimi, S., Vasiliu, A. and Weiffen, B. (2007) 'Negative Results in Social Science.' *European Political Science*, 6(1) pp. 51–68.

Levin, J. R. (1998) 'What If There Were No More Bickering About Statistical Significance Tests?' *Research in the Schools*, 5(2) pp. 43–53.

Levine, T. R., Weber, R., Hullett, C., Park, H. S. and Lindsey, L. L. M. (2008) 'A critical assessment of null hypothesis significance testing in quantitative communication research.' *Human Communication Research*, 34(2) pp. 171–187.

Levitas, R. (2012) *There may be 'trouble' ahead: what we know about those 120,000 'troubled' families*. *Poverty and Social Exclusion in the UK Policy series: Working paper*.

Li, R.-H. and Belford, G. G. (2002) 'Instability of decision tree classification algorithms.' *In 8th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 550–575.

Liaw, A., Wiener, M., Breiman, L. and Cutler, A. (2015) *Package 'randomForest.'*

Lindsay, S. (2015) 'Replication in Psychological Science.' *Psychological Science*, 26(12) pp. 1827–1832.

Liu, S., Maljovec, D., Wang, B., Bremer, P.-T. and Pascucci, V. (2017) 'Visualizing High-Dimensional Data : Advances in the Past Decade.' *IEEE Transactions on Visualization and Computer Graphics*, 23(3) pp. 1249–1268.

Lo, A., Chernoff, H., Zheng, T. and Lo, S.-H. (2015) 'Why significant variables aren't automatically good predictors.' *Proceedings of the National Academy of Sciences*, 112(45) pp. 13892–13897.

Loh, W.-Y. (2014) 'Fifty Years of Classification and Regression Trees.' *International Statistical Review*, 82(2).

Lott, J. R. J. and Whitley, J. (2001) *Abortion and Crime: Unwanted Children and Out-of-Wedlock Births*. *Yale Law & Economics Research Paper No. 254*.

Lovell, M. C. (1983) 'Data Mining.' *The Review of Economics and Statistics*, 65(1) pp. 1–12.

Luskin, R. C. (1991) 'Abusus Non Tollit Usum: Standardized Coefficients, Correlations, and R2s.'

American Journal of Political Science, 35(4) pp. 1032–1046.

Lykken, D. T. (1968) 'Statistical significance in psychological research.' *Psychological bulletin*, 70(3) pp. 151–159.

Ma, Y., Liu, B., Wong, C. K., Yu, P. S. and Lee, S. M. (2000) 'Targeting the right students using data mining.' *In KDD '00 Proceedings of the sixth ACM SIGKDD International conference on Knowledge discovery and data mining*, pp. 457–464.

Van Der Maaten, L. and Hinton, G. (2008) 'Visualizing Data using t-SNE.' *Journal of Machine Learning Research*, 9 pp. 2579–2605.

Macqueen, J. (1967) 'Some Methods For Classification And Analysis of Multivariate Observations.' *In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Manhães, L. M. B., da Cruz, S. M. S. and Zimbrão, G. (2015) 'Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree Programs.' *In Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pp. 247–253.

Mann, A. (2016) 'Core Concepts: Computational social science.' *Proceedings of the National Academy of Sciences*, 113(3) pp. 468–470.

Matejka, J. and Fitzmaurice, G. (2017) 'Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing.' *In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, pp. 1290–1294.

Mazzocchi, F. (2015) 'Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science.' *EMBO reports*, 16(10) pp. 1250–1255.

McAfee, A. and Brynjolfsson, E. (2012) 'Big Data: The management revolution.' *Harvard Buiness Review*, 90(10) pp. 61–68.

McFarland, D. A. and McFarland, H. R. (2015) 'Big Data and the danger of being precisely inaccurate.' *Big Data & Society*, July-Dec(December) pp. 1–4.

McGregor, J. P. (1993) 'Procrustus and the regression model: On the misuse of the regression model.' *PS: Political Science and Politics*, 26(4) pp. 801–804.

McIntyre, N. and Pegg, D. (2018) *Councils use 377,000 people's data in efforts to predict child abuse*. The Guardian. [Online] [Accessed on 18th September 2018] https://www.theguardian.com/society/2018/sep/16/councils-use-377000-peoples-data-in-efforts-to-predict-child-abuse.

Melamed, D., Breiger, R. L. and Schoon, E. (2013) 'The Duality of Clusters and Statistical Interactions.' *Sociological Methods and Research*, 42(1) pp. 41–59.

Milborrow, S. (2013) *Package 'rpart.plot.'*

Milligan, G. W. and Cooper, M. C. (1985) 'An Examination of Procedures for Determining the Number of Clusters in a Data Set.' *Psychometrika*, 50(2) pp. 159–179.

Ministry of Housing Communities & Local Government (2018) *National Evaluation of the Troubled Families Programme 2015-2020: Family Outcomes - national and local datasets, Part 3*.

Mogie, M. (2004) 'In support of null hypothesis significance testing.' *Proceedings: Biological sciences*, 271(3) pp. S82–S84.

Moise, G. and Sander, J. (2008) 'Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering.' *In Conference on Knowledge discovery and data*, pp. 533–541.

Morgan, J. N. and Messenger, R. C. (1973) *THAID A Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables*. Ann Arbor, Michigan: Survey Research Center, Institute for

Social Research, University of Michigan.

Morgan, J. N. and Sonquist, J. A. (1963) 'Problems in the Analysis of Survey Data, and a Proposal.' *Journal of the American Statistical Association*, 58(302) pp. 415–434.

Morrison, D. E. and Henkel, R. E. (eds) (1970) *The Significance Test Controversy: A Reader*. 2nd ed., New Brunswick: Transaction Publishers.

Muchlinski, D., Siroky, D., He, J. and Kocher, M. (2016) 'Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data.' *Political Analysis*, 24(1) pp. 87–103.

Murray, G. R., Riley, C. and Scime, A. (2009) 'Pre-election polling: Identifying likely voters using iterative expert data mining.' *Public Opinion Quarterly*, 73(1) pp. 159–171.

National Institute of Economic and Social Research (2016) *Press Release: No evidence Troubled Families Programme had any significant impact on key objectives, NIESR evaluation finds*. *National Institute of Economic and Social Research*.

National Research Council (2012) *Deterrence and the Death Penalty*. Nagin, D. S. and Pepper, J. V. (eds). Washington, DC: The National Academies Press.

Nelder, J. A. (1999) 'Statistics for the Millennium: From Statistics to Statistical Science.' *Journal of the Royal Statistical Society: Series D (The Statistician)*, 48(2) pp. 257–269.

Nelder, J. A. and Wedderburn, R. W. M. (1972) 'Generalized Linear Models.' *Journal of the Royal Statistical Society. Series A (General)*, 135(3) pp. 370–384.

Nickerson, R. S. (2000) 'Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy.' *Psychological methods*, 5(2) pp. 241–301.

Nielsen, C. B. (2016) 'Visualization: A Mind-Machine Interface for Discovery.' *Trends in Genetics*. Elsevier Ltd, 32(2) pp. 73–75.

Nosek, B. A. (2015) 'Promoting an open research culture.' *Science*, 348(6242) p. 1422.

Nosek, B. A., Spies, J. R. and Motyl, M. (2012) 'Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability.' *Perspectives on Psychological Science*, 7(6) pp. 615–631.

Office for National Statistics (2014) *2011 Census Glossary of Terms*.

Office for National Statistics (2016) *ONS Postcode Directory (August 2016) Version 2*. [Online] [Accessed on 25th September 2016] http://ons.maps.arcgis.com/home/item.html?id=5a656df5f06b4325aa83f907cf0e8d04.

Office for National Statistics (2017) *Output Areas*. [Online] [Accessed on 17th July 2017] https://www.ons.gov.uk/census/2001censusandearlier/dataandproducts/outputgeography/outputareas.

Open Science Collaboration (2015) 'Estimating the reproducibility of psychological science.' *Science*, 349(6251).

Orlitzky, M. (2012) 'How Can Significance Tests Be Deinstitutionalized?' *Organizational Research Methods*, 15(2) pp. 199–228.

Ortega, A. and Navarrete, G. (2017) *Bayesian Hypothesis Testing: An Alternative to Null Hypothesis Significance Testing (NHST) in Psychology and Social Sciences*. *Bayesian Inference*. Intech.

Pearson, K. (1900) 'X. On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling.' *Philosophical Magazine Series 5*, 50(302).

Pearson, K., Yule, G. U., Blanchard, N. and Lee, A. (1903) 'The Law of Ancestral Heredity.' *Biometrika*, 2(2) pp. 211–236.

Pegg, D. and McIntyre, N. (2018) *Child abuse algorithms: from science fiction to cost-cutting reality*. The Guardian. [Online] [Accessed on 18th September 2018] https://www.theguardian.com/society/2018/sep/16/child-abuse-algorithms-from-science-fiction-to-cost-cutting-reality.

Peña-Ayala, A. (2014) 'Educational data mining: A survey and a data mining-based analysis of recent works.' *Expert Systems with Applications*, 41 pp. 1432–1462.

Peng, R. (2015) 'The Reproducibility Crisis in Science.' *Significance*, June pp. 30–32.

Perezgonzalez, J. D. (2015) 'Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing.' *Frontiers in Psychology*, 6(March).

Pezzotti, N., Lelieveldt, B. P. F., Van Der Maaten, L., Höllt, T., Eisemann, E. and Vilanova, A. (2017) 'Approximated and User Steerable tSNE for Progressive Visual Analytics.' *IEEE Transactions on Visualization and Computer Graphics*, 23(7) pp. 1739–1752.

Piscopo, A., Siebes, R. and Hardman, L. (2015) 'Predicting sense of community and participation by applying machine learning to open government data.' *In Data for Policy 2015 - Policy making in the Big Data era: Opportunities & Challenges*. Cambridge.

Platzer, A. (2013) 'Visualization of SNPs with t-SNE.' *PLoS ONE*, 8(2).

Quinlan, J. R. (1986) 'Induction of decision trees.' *Machine Learning*, 1(1) pp. 81–106.

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Rand, W. M. (1971) 'Objective Criteria for the Evaluation of Clustering Methods.' *Journal of the American Statistical Association*, 66(336) pp. 846–850.

Renaud, O. and Victoria-Feser, M. P. (2010) 'A robust coefficient of determination for regression.' *Journal of Statistical Planning and Inference*. Elsevier, 140(7) pp. 1852–1862.

Reyes, J. W. (2007) 'Environmental Policy as Social Policy? The Impact of Childhood Lead Exposure on Crime.' *The B.E. Journal of Economic Analysis & Policy*, 7(1).

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016) '"Why should I trust you?" Explaining the predictions of any classifier.' *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, pp. 1135–1144.

Ridgeway, G. (2017) *Package 'gbm.'*

Rihoux, B. and Ragin, C. C. (eds) (2009) *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*. SAGE Publications, Inc.

Ripley, B. and Venables, W. (2016) *Package 'nnet.'*

Ritschard, G. (2014) 'CHAID and Earlier Supervised Tree Methods.' *In Cotemporary Issues in Exploratory Data Mining in the Behavioural Sciences*. Routledge, pp. 48–76.

Rokach, L. and Maimon, O. (2015) *Data Mining With Decision Trees: Theory and Applications*. 2nd ed., World Scientific Publishing Co. Pte. Ltd.

Romero, C. and Ventura, S. (2007) 'Educational data mining: A survey from 1995 to 2005.' *Expert Systems with Applications*, 33(1) pp. 135–146.

Rosenthal, R. (1979) 'The "File Drawer Problem" and Tolerance for Null Results.' *Psychological Bulletin*, 86(3) pp. 638–641.

Rousseeuw, P. J. (1987) 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.' *Journal of Computational and Applied Mathematics*, 20 pp. 53–65.

Rozeboom, W. W. (1960) 'The fallacy of the null-hypothesis significance test.' *Psychological Bulletin*, 57(5) pp. 416–428.

Ruger, T. W., Kim, P. T., Martin, A. D. and Quinn, K. (2004) 'The Supreme Court Forcasting Project:

Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking.' *Columbia Law Review*, 104 pp. 1150–1210.

Sadler, K., Akister, J. and Burch, S. (2015) 'Who are the young people who are not in education, employment or training? An application of the risk factors to a rural area in the UK.' *International Social Work*, 58(4) pp. 508–520.

Salazar, A., Gosalbez, J., Bosch, I., Miralles, R. and Vergara, L. (2004) 'A case study of knowledge discovery on academic achievement, student desertion and student retention.' *In ITRE 2004: 2nd International Conference Information Technology: Research and Education, Proceedings*, pp. 150–154.

Savage, M. and Burrows, R. (2007) 'The Coming Crisis of Empirical Sociology.' *Sociology*, 41(5) pp. 885–899.

Schmidt, F. L. and Hunter, J. (1997) 'Eight common but false objections to the discontinuation of significance testing in the analysis of research data.' *In* Harlow, L. L., Mulaik, S. A., and Steiger, J. H. (eds) *What if There Were no Significance Tests?* London: Routledge.

Schmidt, S. (2009) 'Shall we really do it again? The powerful concept of replication is neglected in the social sciences.' *Review of General Psychology*, 13(2) pp. 90–100.

Van de Schoot, R., Hoijtink, H. and Jan-Willem, R. (2011) 'Moving beyond traditional null hypothesis testing: Evaluating expectations directly.' *Frontiers in Psychology*, 2(FEB) pp. 1–5.

Scime, A. and Murray, G. R. (2013) 'Social Science Data Analysis: The Ethical Imperative.' *In Ethical Data Mining Applications for Socio-Economic Development*. IGI Global, pp. 131–147.

Sedlmeier, P. and Gigerenzer, G. (1989) 'Do Studies of Statistical Power Have an Effect on the Power of Studies?' *Psychological Bulletin* pp. 309–316.

Sembiring, R., Zain, J. and Embong, A. (2010) 'Clustering high dimensional data using subspace and projected clustering algorithms.' *International Journal of Computer Science and Information Technology*, 2(4) pp. 162–170.

Shaffer, J. P. (1995) 'Multiple Hypothesis Testing.' *Annual Review of Psychology*, 46 pp. 561–584.

Shah, D. V., Cappella, J. N. and Neuman, W. R. (2015) 'Big Data, Digital Media, and Computational Social Science.' Shah, D. V., Cappella, J. N., and Neuman, W. R. (eds) *The ANNALS of the American Academy of Political and Social Science*, 659(1) pp. 6–13.

Shildrick, T., MacDonald, R. and Furlong, A. (2016) 'Not single spies but in battalions: a critical, sociological engagement with the idea of so-called "Troubled Families."' *Sociological Review*, 64(4) pp. 821–836.

Shmueli, G. (2010) 'To explain or to predict?' *Statistical Science*, 25(3) pp. 289–310.

Shoresh, N. and Wong, B. (2012) 'Points of view: Data exploration.' *Nature Methods*. Nature Publishing Group, 9(1) p. 5.

Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011) 'False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.' *Psychological Science*, 22(11) pp. 1359–1366.

Simons, D. J. (2014) 'The Value of Direct Replication.' *Perspectives on Psychological Science*, 9(1) pp. 76–80.

Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2014) 'P-curve: A key to the File-Drawer.' *Journal of Experimental Psychology: General*, 143(2) pp. 534–547.

Singer, N. (2012a) *F.T.C Opens an Inquiry Into Data Brokers*. The New York Times. [Online] [Accessed on 28th February 2017] http://www.nytimes.com/2012/12/19/technology/ftc-opens-an-inquiry-into-data-brokers.html.

Singer, N. (2012b) *Mapping, and Sharing the Consumer Genome*. The New York Times. [Online] [Accessed on 28th February 2017] http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html.

Smith, R. A., Levine, T. R., Lachlan, K. A. and Fediuk, T. A. (2002) 'The High Cost of Complexity in Experimental Design and Data Analysis Type I and Type II Error Rates in Multiway ANOVA.' *Human Communication Research*, 28(4) pp. 515–530.

Sonquist, J. A. and Morgan, J. N. (1964) *The Detection of Interaction Effects: A Report on a Computer Program for the Selection of Optional Combinations of Explanatory Variables*. Survey Research Center, Institute for Social Research, University of Michigan.

Sorenson, H. W. (1970) 'Least-squares estimation: from Gauss to Kalman.' *IEEE Spectrum*, 7 pp. 63–68.

Soyer, E. and Hogarth, R. M. (2012) 'The illusion of predictability: How regression statistics mislead experts.' *International Journal of Forecasting*, 28(3) pp. 695–711.

Stambuk, A., Stambuk, N. and Konjevoda, P. (2007) 'Application of Kohonen Self-Organizing Maps (SOM) Based Clustering for the Assessment of Religious Motivation.' *In Proceedings of the ITI 29th International Conference on Information Technology Interfaces*. Cavtat, Croatia, pp. 87–91.

Stanton, J. M. (2001) 'Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors.' *Journal of Statistics Education*, 9(3).

Steel, E. (2014) *Data Brokers Start to Feel the Net Tighten*. Financial Times. [Online] [Accessed on 28th February 2017] https://www.ft.com/content/1fdc6e0c-df60-11e3-86a4-00144feabdc0.

Steinbach, M., Ertoz, L., Kumar, V., Ertöz, L. and Kumar, V. (2004) 'The challenges of clustering high dimensional data.' *New Directions in Statistical Physics* pp. 1–33.

Strobl, C., Boulesteix, A.-L. and Augustin, T. (2007) 'Unbiased Split Selection for Classification Trees Based on the Gini Index.' *Computational Statistics & Data Analysis*, 52(1).

Swinford, S. (2016) *Ministers accused of suppressing report which found troubled families programme has had 'no discernible impact.'* The Telegraph. [Online] [Accessed on 11th August 2016] http://www.telegraph.co.uk/news/2016/08/08/ministers-accused-of-suppressing-report-which-found-troubled-fam/.

Szucs, D. (2016) 'A Tutorial on Hunting Statistical Significance by Chasing N.' *Frontiers in Psychology*, 7(September) pp. 1–10.

Szucs, D. and Ioannidis, J. P. A. (2017) 'When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment.' *Frontiers in Human Neuroscience*, 11(August).

Taagepera, R. (2008) *Making Social Sciences More Scientific: The Need for Predictive Models*. OUP Oxford.

The British Psychological Society (2017) *Behaviour Change: School attendance, exclusion and persistent absence*.

Therneau, T., Atkinson, B. and Ripley, B. (2017) *Package 'rpart.'*

Therneau, T. M. and Atkinson, E. J. (2015) *An Introduction to Recursive Partitioning Using the RPART Routines*.

Thielman, S. (2015) 'Experian hack raises doubts about security of credit database, advocates say.' *The Guardian*. [Online] 8th October. https://www.theguardian.com/business/2015/oct/08/experian-hack-advocates-question-security-database.

Thomas, E. H. and Galambos, N. (2004) 'What satisfies students? Mining students-opinion data with regression and decision tree analysis.' *Research in Higher Education*, 45(3) pp. 251–269.

Thompson, B. (1993) 'The Use of Statistical Significance Tests in Research: Bootstrap and Other Alternatives.' *The Journal of Experimental Education*, 61(4) pp. 361–377.

Trafimow, D. and Marks, M. (2015) 'Editorial.' *Basic and Applied Social Psychology*, 37(1) pp. 1–2.

Tufte, E. R. (1969) 'Improving Data Analysis in Political Science.' *World Politics*, 21(4) pp. 641–654.

Tukey, J. W. (1962) 'The Future of Data Analysis.' *The Annals of Mathematical Statistics*, 33(1) pp. 1–67.

Tukey, J. W. (1977) *Exploratory Data Analysis*. Addison-Wesley Publishing Company.

Tukey, J. W. (1991) 'The Philosophy of Multiple Comparisons.' *Statistical Science*, 6(1) pp. 100–116.

UK Data Service (2011) *Office for National Statistics: 2011 Census Aggregate Data*. [Online] [Accessed on 24th October 2016] http://infuse.ukdataservice.ac.uk/.

Uprichard, E. (2009) 'Introducing Cluster Analysis: What Can It Teach Us about the Case?' *In* Byrne, D. and Ragin, C. C. (eds) *The Sage Handbook of Case-Based Methods*. Sage, pp. 132–147.

Vatcheva, K., Lee, M., McCormick, J. and Rahbar, M. (2016) 'The Effect of Ignoring Statistical Interactions in Regression Analyses Conducted in Epidemiologic Studies: An Example with Survival Analysis Using Cox Proportional Hazards Regression Model.' *Epidemiology: Open Access*, 06(01) pp. 1–7.

Veltri, G. A. (2017) 'Big Data is not only about data: The two cultures of modelling.' *Big Data & Society*, January-Ju.

Vidgen, B. and Yasseri, T. (2016) 'P-Values: Misunderstood and Misused.' *Frontiers in Physics*, 4(March) pp. 10–14.

Ward, M. D., Greenhill, B. D. and Bakke, K. M. (2010) 'The perils of policy by p-value: Predicting civil conflicts.' *Journal of Peace Research*, 47(4) pp. 363–375.

Wasserstein, R. L. and Lazar, N. A. (2016) 'The ASA's Statement on p-Values: Context, Process, and Purpose.' *The American Statistician*. Taylor & Francis, 70(2) pp. 129–133.

Wattenberg, M., Viégas, F. and Johnson, I. (2016) *How to Use t-SNE Effectively*. Distill. [Online] [Accessed on 8th August 2017] http://distill.pub/2016/misread-tsne.

Watts, D. J. (2013) 'Computational Social Science: Exciting Progress and Future Directions.' *The Bridge*, 43(4) pp. 5–10.

Web of Science (2017) *Web of Science*. [Online] [Accessed on 28th February 2017] http://apps.webofknowledge.com/summary.do?SID=4Ep4D5ISdBiHokoItfs&product=WOS&parentQid=2&qid=3&search_mode=GeneralSearch&colName=WOS&mode=refine.

Weerts, D. J. and Ronca, J. M. (2009) 'Using classification trees to predict alumni giving for higher education.' *Education Economics*, 17(1) pp. 95–122.

Welch, I. and Goyal, A. (2007) 'A comprehensive look at the empirical performance of equity premium prediction.' *The Review of Financial Studies*, 21(4) pp. 1455–1508.

Welles, B. F. (2014) 'On minorities and outliers: The case for making Big Data small.' *Big Data & Society*, 1(1).

Wenham, A. (2017) 'Struggles and Silences: Young People and the "Troubled Families Programme."' *Social Policy and Society*, 16(1) pp. 143–153.

Wilcox, R. R. (1998) 'How many discoveries have been lost by ignoring modern statistical methods?' *American Psychologist*, 53(3) pp. 300–314.

Wilkinson, L. and the Taskforce on Statistical Inference (1999) 'Statistical Methods in Psychology Journals: Guidelines and Explanations.' *American Psychologist* pp. 594–604.

Wills, J., Whittaker, A., Rickard, W. and Felix, C. (2017) 'Troubled, Troubling or in Trouble: The Stories of "Troubled Families."' *British Journal of Social Work*, 47(4) pp. 989–1006.

Witten, I. H., Frank, E. and Hall, M. A. (2011) *Data Mining Practical Machine Learning Tools and Techniques*. 3rd ed., Morgan Kaufmann.

Woodside, A. G. (2016) 'The good practices manifesto: Overcoming bad practices pervasive in current research in business.' *Journal of Business Research*. Elsevier Inc., 69(2) pp. 365–381.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J. and Steinberg, D. (2007) 'Top 10 algorithms in data mining.' *Knowledge and Information Systems*, 14(1) pp. 1–37.

Xantura (2018a) *Xantura :: Insight: Delivering shared intelligence... where and when it matters most*. Xantura. [Online] [Accessed on 18th September 2018] https://www.xantura.com/products/insight.

Xantura (2018b) *Xantura :: Troubled families*. Xantura. [Online] [Accessed on 18th September 2018] https://www.xantura.com/focus-areas/troubled-families.

Xie, Y. (2013) 'Population heterogeneity and causal inference.' *Proceedings of the National Academy of Sciences of the United States of America*, 110(16) pp. 6262–8.

Yang, Q. and Wu, X. (2006) '10 Challenging Problems in Data Mining Research.' *IEEE International Conference on Data Mining*, 5(4) pp. 597–604.

Yarkoni, T. and Westfall, J. (2017) 'Choosing prediction over explanation in psychology: Lessons from machine learning.' *Perspectives on Psychological Science*, 12(6) pp. 1100–1122.

Yule, G. U. (1897) 'On the Theory of Correlation.' *Journaof the Royal Statistical Society*, 60(4) pp. 812–854.

Ziliak, S. T. (2012) 'Visualizing uncertainty: On Soyer's and Hogarth's "The illusion of predictability: How regression statistics mislead experts."' *International Journal of Forecasting*. Elsevier B.V., 28(3) pp. 712–714.

Ziliak, S. T. and McCloskey, D. N. (2009) *The Cult of Statistical Significance*. *Section on Statistical Education - JSM 2009*. Washington, DC.

Zuur, A. F., Ieno, E. N. and Elphick, C. S. (2010) 'A protocol for data exploration to avoid common statistical problems.' *Methods in Ecology and Evolution*, 1(1) pp. 3–14.

APPENDICES

APPENDIX A

A1: Attributes utilised as predictors for the models predicting cluster assignment from place-based data

This section details the set of attributes that were utilised as predictors in section 6.3.4.1, which used 'place-based' attributes to predict cluster assignment. Data was obtained from: the 2011 Census data (UK Data Service, 2011), and linked to each family via the Output Area (OA) they lived in at the start of intervention; and Police crime data (Home Office, 2016) for the year 2011, and linked to each family via the LSOA area they lived in. Whilst experiments were performed using a larger set of attributes (for instance, including all religion and all ethnic group attributes), these were reduced down to a smaller set since the inclusion of all attributes did not increase accuracy, and in order to potentially obtain a more understandable model. The smaller set of 22 attributes included the most populous values; for instance, the 'white' ethnic group, was the largest overall so that was included, but in some areas the 'Asian/Asian British' population was larger, so this too was included. The table details the attributes utilised in the final models:

Attribute Description	Source
Population Density of OA	Census, 2011
Ethnic Group – Percentage 'white' people in OA	Census, 2011
Ethnic Group – Percentage 'Asian/Asian British' people in OA	Census, 2011
Household Deprivation – Percentage households not deprived in OA	Census, 2011
Place of Birth – Percentage of people born in the UK in OA	Census, 2011
Place of Birth – Percentage of people born in Asia/Middle East in OA	Census, 2011
Qualifications – Percentage of people with no qualifications in OA	Census, 2011
Religion – Percentage of people who are Christian in OA	Census, 2011
Religion – Percentage of people who are Muslim in OA	Census, 2011
Religion – Percentage of people who have no religion in OA	Census, 2011
Tenure – Percentage of households that are owned in OA	Census, 2011
Tenure – Percentage of households that are privately rented in OA	Census, 2011
Tenure – Percentage of households that are social rented in OA	Census, 2011
Economic Activity – Percentage of economically active people in OA	Census, 2011
Language – Percentage of households in OA where English is first language of all adults	Census, 2011
Household Size – Percentage of single-person households in OA	Census, 2011
Household Composition – Percentage of lone-parent households in OA	Census, 2011
General Health – Percentage of people in OA with 'bad' or 'very bad' general health	Census, 2011
Long-term Health – Percentage of people in OA with limited long-term health	Census, 2011
Police – Anti-social behaviour – Percentage of crimes in the LSOA that were anti-social behaviour (as a percentage of total number crimes in LSOA), 2011	Police Data, 2011
Police – Burglary – Percentage of crimes in the LSOA that were burglary (as a percentage of total number crimes in LSOA), 2011	Police Data, 2011
Police – Violent crime – Percentage of crimes in the LSOA that were violent (as a percentage of total number crimes in LSOA), 2011	Police Data, 2011

A2: Full Variable Importance scores for each model, with model details

Decision tree

The model had a test set accuracy of 28.17%, which was almost the same as the baseline accuracy (28.02%). The tree was pruned with a CP value of 0.0033 and utilised equal weights. The table contains the full Variable Importance scores:

Attribute	Variable Importance Score
Tenure – Percentage of households that are privately rented in OA	19
Long-term Health – Percentage of people in OA with limited long-term health	13
General Health – Percentage of people in OA with 'bad' or 'very bad' general health	12
Qualifications – Percentage of people with no qualifications in OA	11
Tenure – Percentage of households that are social rented in OA	11
Household Size – Percentage of single-person households in OA	10
Economic Activity – Percentage of economically active people in OA	7
Place of Birth – Percentage of people born in the UK in OA	7
Household Composition – Percentage of lone-parent households in OA	3
Household Deprivation – Percentage households not deprived in OA	2
Religion – Percentage of people who have no religion in OA	2
Religion – Percentage of people who are Muslim in OA	1
Language – Percentage of households in OA where English is first language of all adults	1

Random forest

The model had a test set accuracy of 19.06%, which was lower than the baseline accuracy (28.02%). The forest contained 1150 trees, and tried 4 attributes at each split. The full Variable Importance scores are contained in the table, in order and in terms of mean decrease in accuracy:

Attribute	Variable Importance Score
Long-term Health – Percentage of people in OA with limited long-term health	9.04
Economic Activity – Percentage of economically active people in OA	7.50
Place of Birth – Percentage of people born in the UK in OA	6.83
Household Size – Percentage of single-person households in OA	6.63
Qualifications – Percentage of people with no qualifications in OA	6.42
Ethnic Group – Percentage 'Asian/Asian British' people in OA	5.72
Household Deprivation – Percentage households not deprived in OA	5.64
Place of Birth – Percentage of people born in Asia/Middle East in OA	5.42
General Health – Percentage of people in OA with 'bad' or 'very bad' general health	5.33
Religion – Percentage of people who are Christian in OA	5.27
Tenure – Percentage of households that are privately rented in OA	5.25
Ethnic Group – Percentage 'white' people in OA	4.77
Household Composition – Percentage of lone-parent households in OA	4.22
Tenure – Percentage of households that are social rented in OA	4.13
Religion – Percentage of people who are Muslim in OA	3.82
Language – Percentage of households in OA where English is first language of all adults	3.53
Police – Violent crime – Percentage of crimes in the LSOA that were violent (as a percentage of total number crimes in LSOA), 2011	2.51
Population Density of OA	2.25

Police – Anti-social behaviour – Percentage of crimes in the LSOA that were anti-social behaviour (as a percentage of total number crimes in LSOA), 2011	2.11
Police – Burglary – Percentage of crimes in the LSOA that were burglary (as a percentage of total number crimes in LSOA), 2011	1.04
Tenure – Percentage of households that are owned in OA	1.00
Religion – Percentage of people who have no religion in OA	0.48

Generalized boosted model

The model had a test set accuracy of 26.78%, which was lower than the baseline accuracy (28.02%). A model with 3000 trees was built, with shrinkage value of 0.001, and depth of 1; the best cross-validation iteration was 2275. The full Variable Importance scores are in the table:

Attribute	Variable Importance Score
Tenure – Percentage of households that are privately rented in OA	11.38
Long-term Health – Percentage of people in OA with limited long-term health	8.47
Household Size – Percentage of single-person households in OA	6.85
Economic Activity – Percentage of economically active people in OA	6.52
General Health – Percentage of people in OA with 'bad' or 'very bad' general health	6.18
Ethnic Group – Percentage 'white' people in OA	5.14
Religion – Percentage of people who have no religion in OA	4.93
Population Density of OA	4.86
Tenure – Percentage of households that are social rented in OA	4.86
Tenure – Percentage of households that are owned in OA	4.40
Police – Violent crime – Percentage of crimes in the LSOA that were violent (as a percentage of total number crimes in LSOA), 2011	4.20
Police – Anti-social behaviour – Percentage of crimes in the LSOA that were anti-social behaviour (as a percentage of total number crimes in LSOA), 2011	3.75
Police – Burglary – Percentage of crimes in the LSOA that were burglary (as a percentage of total number crimes in LSOA), 2011	3.60
Religion – Percentage of people who are Muslim in OA	3.38
Place of Birth – Percentage of people born in Asia/Middle East in OA	3.28
Language – Percentage of households in OA where English is first language of all adults	3.12
Ethnic Group – Percentage 'Asian/Asian British' people in OA	3.12
Place of Birth – Percentage of people born in the UK in OA	2.87
Household Deprivation – Percentage households not deprived in OA	2.52
Qualifications – Percentage of people with no qualifications in OA	2.36
Religion – Percentage of people who are Christian in OA	2.14
Household Composition – Percentage of lone-parent households in OA	2.08

Multinomial logistic regression

The model had a test set accuracy of 26.32%, which was lower than the baseline accuracy (28.02%). The reference level for the target attribute was set to cluster 11, as this was the largest group and contained families with no events, so could be thought of as a baseline to some degree. The residual deviance was 5972.846, Akaike Information Criterion was 6432.846, and the effective degrees of freedom were 230. The table contains the intercept coefficients together with their standard errors in brackets. Values that were significant at the $p < 0.1$ (*), $p < 0.05$ (**) and $p < 0.01$ (***) are labelled.

Predictors:	Target Attribute (cluster assignment):									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Population Density of OA	-0.416 (0.826)	0.147 (0.745)	1.914* (1.031)	0.371 (1.376)	-1.275 (2.909)	0.892 (1.530)	1.345 (2.242)	0.923 (0.866)	0.432 (0.819)	0.394 (0.947)
Ethnic Group – Percentage 'white' people in OA	0.504 (1.752)	-1.748 (1.646)	-0.661 (2.479)	0.355 (3.486)	5.516 (5.227)	-4.924 (3.594)	-1.651 (4.897)	-3.587* (1.943)	-0.274 (1.822)	0.234 (2.066)
Ethnic Group – Percentage 'Asian/Asian British' people in OA	-2.430 (2.859)	-3.261 (2.685)	0.916 (3.691)	-2.030 (4.904)	2.369 (9.909)	-6.066 (5.659)	4.420 (8.611)	-4.004 (3.118)	-2.360 (2.897)	3.724 (3.420)
Household Deprivation – Percentage households not deprived in OA	3.282 (1.996)	2.723 (1.904)	4.741* (2.735)	1.410 (3.954)	5.521 (6.575)	8.126** (3.931)	1.008 (5.918)	-0.414 (2.226)	1.760 (2.081)	-0.548 (2.339)
Place of Birth – Percentage of people born in the UK in OA	-1.282 (2.815)	1.919 (2.698)	-2.854 (3.924)	-1.478 (5.481)	-15.889* (8.463)	-0.232 (5.717)	-2.913 (8.471)	4.780 (3.223)	-0.926 (2.904)	-1.789 (3.279)
Place of Birth – Percentage of people born in Asia/Middle East in OA	5.105 (4.415)	6.731 (4.104)	-5.465 (5.852)	5.502 (7.801)	2.428 (14.146)	19.160** (8.903)	2.217 (12.488)	9.106* (4.867)	2.292 (4.436)	-6.112 (5.414)
Qualifications – Percentage of people with no qualifications in OA	2.641 (2.007)	2.241 (1.911)	2.053 (2.848)	4.213 (3.821)	-5.976 (6.061)	7.161* (3.870)	1.881 (5.579)	1.567 (2.173)	1.544 (2.156)	-0.728 (2.356)
Religion – Percentage of people who are Christian in OA	1.636 (3.505)	0.227 (3.326)	-10.400** (4.716)	-0.498 (6.902)	8.088 (11.472)	-4.213 (6.702)	0.968 (9.725)	-4.228 (3.674)	-5.983* (3.595)	-6.763* (3.941)
Religion – Percentage of people who are Muslim in OA	1.460 (3.647)	-2.662 (3.438)	-8.223* (4.898)	3.021 (6.920)	9.266 (11.569)	-2.903 (6.757)	-2.900 (9.809)	-6.424* (3.860)	-7.183* (3.704)	-5.923 (4.092)
Religion – Percentage of people who have no religion in OA	4.369 (3.850)	0.855 (3.663)	-7.264 (5.159)	3.382 (7.440)	4.833 (12.929)	-3.934 (7.401)	0.990 (10.904)	-2.661 (4.005)	-6.130 (3.981)	-4.512 (4.310)
Tenure – Percentage of households that are owned in OA	-1.700 (4.171)	0.012 (4.147)	16.105* (8.290)	3.628 (8.749)	-8.032 (11.583)	-4.452 (8.025)	-11.657 (9.067)	4.450 (5.252)	-4.746 (4.209)	-4.015 (4.847)
Tenure – Percentage of households that are social rented in OA	-2.236 (4.126)	-0.329 (4.103)	14.877* (8.216)	0.952 (8.516)	-3.956 (11.333)	-2.842 (7.903)	-11.043 (8.805)	3.931 (5.188)	-6.087 (4.175)	-4.099 (4.780)
Tenure – Percentage of households that are privately rented in OA	-2.459 (4.187)	1.325 (4.154)	15.907* (8.329)	0.687 (8.578)	-6.184 (11.585)	-1.408 (8.025)	-13.139 (9.239)	4.404 (5.270)	-5.429 (4.240)	-3.327 (4.863)

Economic Activity – Percentage of economically active people in OA	0.291 (1.876)	-0.825 (1.771)	0.630 (2.627)	-0.161 (3.531)	-12.116** (5.944)	-0.923 (3.764)	2.025 (5.566)	1.836 (2.055)	0.568 (1.978)	-0.200 (2.190)
Language – Percentage of households in OA where English is first language of all adults	-2.120 (2.607)	-2.750 (2.459)	-0.026 (3.453)	0.313 (4.394)	16.140* (8.858)	14.040*** (5.414)	7.121 (7.978)	0.031 (2.971)	-2.015 (2.658)	2.435 (3.106)
Household Size – Percentage of single-person households in OA	-2.566** (1.289)	-2.881** (1.225)	0.720 (1.786)	0.085 (2.570)	-3.047 (3.715)	-3.410 (2.603)	1.579 (3.607)	-2.251 (1.410)	-1.657 (1.356)	-2.243 (1.500)
Household Composition – Percentage of lone-parent households in OA	-0.885 (2.382)	-0.178 (2.227)	4.070 (3.376)	-0.505 (4.970)	-1.989 (7.000)	-2.222 (4.618)	3.267 (6.868)	-1.056 (2.549)	-0.209 (2.521)	-0.786 (2.734)
General Health – Percentage of people in OA with 'bad' or 'very bad' general health	-1.195 (5.107)	3.918 (5.059)	0.216 (7.710)	11.689 (8.888)	1.412 (16.797)	12.393 (10.616)	-0.247 (15.368)	-0.459 (5.767)	11.600** (5.691)	-1.614 (6.348)
Long-term Health – Percentage of people in OA with limited long-term health	6.786* (3.497)	0.911 (3.528)	4.961 (5.123)	1.382 (6.332)	-12.138 (12.347)	-4.339 (7.533)	-4.497 (11.246)	3.485 (3.891)	-4.260 (4.062)	2.810 (4.242)
Police – Anti-social Behaviour – Percentage of crimes in the LSOA that were anti-social behaviour (as a percentage of total number crimes in LSOA), 2011	-0.231 (1.482)	0.531 (1.464)	-0.521 (2.191)	-2.573 (2.587)	-2.891 (4.520)	1.445 (3.069)	-3.799 (4.326)	-1.778 (1.634)	-0.031 (1.610)	-0.647 (1.749)
Police – Violent Crime – Percentage of crimes in the LSOA that were violent crime (as a percentage of total number crimes in LSOA), 2011	-3.913 (2.571)	-0.221 (2.461)	-1.435 (3.660)	-2.715 (4.600)	-4.542 (8.156)	-3.539 (5.386)	-2.484 (7.425)	-0.331 (2.738)	-5.772** (2.803)	-4.121 (3.022)
Police – Burglary – Percentage of crimes in the LSOA that were burglary (as a percentage of total number crimes in LSOA), 2011	-0.656 (2.375)	0.801 (2.290)	-2.870 (3.462)	-12.099** (5.198)	-1.099 (7.590)	3.695 (4.790)	3.679 (6.229)	-2.018 (2.629)	1.771 (2.471)	-2.393 (2.813)
Constant	-0.089 (5.846)	0.313 (5.639)	-9.503 (9.640)	-5.375 (11.724)	3.339 (17.319)	-8.448 (11.370)	3.357 (14.822)	-3.446 (6.796)	12.257** (6.007)	9.202 (6.722)

significance levels: $p < 0.1$ (*), $p < 0.05$ (**) and $p < 0.01$ (***)

A3: Simpler Multinomial Logistic Regression Model

The top five most important predictors for the three machine learning models (in Table 23) were collated to produce a list of 8 predictors in order to produce a simpler logistic regression model. Whilst these predictors are still generally correlated (as indicated in Figure 10), it was felt that a smaller set might provide a more understandable model. Utilising the machine learning methods as a form of feature selection provided a data-driven method of feature reduction. The predictors were:

Attribute	Source
Place of Birth – Percentage of people born in the UK in OA	Census, 2011
Qualifications – Percentage of people with no qualifications in OA	Census, 2011
Tenure – Percentage of households that are privately rented in OA	Census, 2011
Tenure – Percentage of households that are social rented in OA	Census, 2011
Economic Activity – Percentage of economically active people in OA	Census, 2011
Household Size – Percentage of single-person households in OA	Census, 2011
General Health – Percentage of people in OA with 'bad' or 'very bad' general health	Census, 2011
Long-term Health – Percentage of people in OA with limited long-term health	Census, 2011

The resulting multinomial logistic regression model had a test set accuracy of 27.86%, which was an improvement of 1.54% over the first model (which had accuracy of 26.32%). This was still lower than the baseline accuracy (28.02%). The reference level for the target attribute was set to cluster 11, as this was the largest group and contained families with no events, so could be thought of as a baseline to some degree. The residual deviance was 6097.708, Akaike Information Criterion was 6277.708, and the effective degrees of freedom were 90. The table contains the intercept coefficients together with their standard errors in brackets. Values that were significant at the $p < 0.1$ (*), $p < 0.05$ (**) and $p < 0.01$ (***) are labelled.

	Target Attribute (cluster assignment):									
Predictors:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Place of Birth – Percentage of people born in the UK in OA	-1.042 (0.886)	-0.217 (0.815)	-1.802 (1.152)	-5.144*** (1.633)	0.835 (2.379)	-1.680 (1.675)	-0.816 (2.467)	0.202 (0.979)	-1.239 (0.890)	0.329 (1.012)
Qualifications – Percentage of people with no qualifications in OA	0.044 (1.667)	0.818 (1.551)	-0.436 (2.305)	2.237 (3.292)	-6.287 (4.746)	2.035 (3.300)	2.061 (4.692)	0.496 (1.812)	0.354 (1.738)	-1.512 (1.908)
Tenure – Percentage of households that are social rented in OA	-0.535 (0.673)	-0.085 (0.643)	-0.129 (0.950)	-2.483* (1.361)	1.133 (2.118)	-0.282 (1.394)	-0.066 (1.925)	0.273 (0.745)	-1.338* (0.699)	0.046 (0.781)
Tenure – Percentage of households that are privately rented in OA	-0.452 (1.037)	1.685* (0.939)	0.482 (1.360)	-3.643* (2.145)	0.116 (2.972)	0.013 (2.049)	-1.855 (2.991)	0.477 (1.128)	-0.179 (1.020)	1.067 (1.146)
Economic Activity – Percentage of economically active people in OA	1.319 (1.724)	0.308 (1.637)	1.470 (2.432)	-0.031 (3.235)	-9.136* (5.316)	1.254 (3.448)	1.635 (5.077)	2.108 (1.902)	0.838 (1.817)	-0.694 (2.018)
Household Size – Percentage of single-person households in OA	-1.769* (1.013)	-2.513** (0.989)	-0.439 (1.409)	0.364 (1.831)	-0.270 (2.963)	-0.952 (2.009)	1.385 (2.738)	-1.747 (1.119)	-1.138 (1.094)	-1.731 (1.202)
General Health – Percentage of people in OA with 'bad' or 'very bad' general health	-0.805 (5.041)	3.875 (5.027)	2.914 (7.627)	12.628 (8.396)	-0.116 (15.927)	11.920 (10.069)	-3.321 (15.116)	-0.872 (5.656)	11.123** (5.626)	-2.258 (6.189)
Long-term Health – Percentage of people in OA with limited long-term health	5.822* (3.231)	-0.323 (3.289)	-0.559 (5.040)	0.301 (5.715)	-6.891 (11.045)	-5.686 (7.009)	-4.227 (10.169)	3.339 (3.641)	-4.574 (3.808)	3.337 (3.979)
Constant	-1.015 (1.740)	-0.632 (1.689)	-1.103 (2.479)	1.233 (3.133)	4.553 (5.507)	-2.010 (3.499)	-3.114 (5.125)	-2.885 (1.956)	0.387 (1.849)	-0.687 (2.081)

Significance levels: $p < 0.1$ (*), $p < 0.05$ (**) and $p < 0.01$ (***)

APPENDIX B

B1: Attributes utilised as predictors for the machine learning models

This section details the full set of attributes that were utilised as predictors in section 7.3.7. In general, three datasets were used: set A, which consisted of all the attributes in the table (below); a smaller set (B), and an even smaller set where the clustering attributes were replaced by the cluster assignment (C). The larger set (A) was utilised in order to make no assumptions about the data and allow a wide variety of attributes to be considered as predictors. However, a more parsimonious set (B) was also utilised to see if that made any difference; this excluded some of the more repetitive attributes. The smaller set of attributes included the most populous values; for instance, the 'white' ethnic group, was the largest overall so that was included, but not the attributes detailing all the other ethnic groups. And substituting the cluster assignment for the clustering attributes (for example, school absence, CIN events, etc.) meant an even simpler dataset (set C). However, dataset C was not utilised for the cluster-level models, as in this case the cluster assignment was irrelevant.

Data from the ECC database and pertaining directly to the family was included, together with attributes that were linked to the area that a family lived in at the start of intervention treatment. The Census 2011 data (UK Data Service, 2011) linked via the family's Output Area code, and the Police data (Home Office, 2016) linked via the family's LSOA code. All attributes, together with which dataset they were in, are detailed below:

Description	Type	Source	Set
Intervention type	Categorical	ECC	A, B, C
Number of people in the family	Integer	ECC	A, B, C
Number of females in the family	Integer	ECC	A
Number of males in the family	Integer	ECC	A
Number of children in the family	Integer	ECC	A, B, C
Number of adults in the family	Integer	ECC	A, B, C
Percentage of unauthorised absence for the previous year, for the family	Count	ECC	A, B
Number of school exclusions for the previous year, for the family	Integer	ECC	A, B
Number of criminal offences committed by adults for the previous year, for the family	Integer	ECC	A, B
Number of criminal offences committed by children for the previous year, for the family	Integer	ECC	A, B
Family had one or more NEET members in the previous year	Binary	ECC	A, B
Family had one or more CIN events in the previous year	Binary	ECC	A, B
Family had one or more CPPs in the previous year	Binary	ECC	A, B
Family had one or more LAC events in the previous year	Binary	ECC	A, B
Family had one or more events classed as domestic abuse in the previous year	Binary	ECC	A
Family had one or more drug/alcohol events in the previous year	Binary	ECC	A

Number of changes of address for the previous year, for the family	Integer	ECC	A
Cluster assignment	Categorical	Derived	A, C
Population Density of OA	Numerical	Census, 2011	A
Ethnic Group – Percentage 'white' people in OA	Numerical	Census, 2011	A, B, C
Ethnic Group – Percentage 'Asian/Asian British' people in OA	Numerical	Census, 2011	A
Ethnic Group – Percentage 'Black/African/Caribbean/Black British' people in OA	Numerical	Census, 2011	A
Ethnic Group – Percentage 'Mixed/Multi Ethnic group' people in OA	Numerical	Census, 2011	A
Ethnic Group – Percentage 'Other' people in OA	Numerical	Census, 2011	A
Percentage of households in OA not deprived	Numerical	Census, 2011	A, B, C
Percentage of people born in the UK in OA	Numerical	Census, 2011	A, B, C
Percentage of people born in the Europe in OA	Numerical	Census, 2011	A
Percentage of people born in the Africa in OA	Numerical	Census, 2011	A
Percentage of people born in the Middle East and Asia in OA	Numerical	Census, 2011	A
Percentage of people born in the Americas and Caribbean in OA	Numerical	Census, 2011	A
Percentage of people born in the Antarctica and Oceania in OA	Numerical	Census, 2011	A
Percentage of people born in Other place in OA	Numerical	Census, 2011	A
Percentage of people with no qualifications in OA	Numerical	Census, 2011	A, B, C
Religion – Percentage of people who are Christian in OA	Numerical	Census, 2011	A, B, C
Religion – Percentage of people who are Muslim in OA	Numerical	Census, 2011	A
Religion – Percentage of people who have no religion in OA	Numerical	Census, 2011	A
Tenure – Percentage of households that are owned in OA	Numerical	Census, 2011	A, B, C
Tenure – Percentage of households that are privately rented in OA	Numerical	Census, 2011	A
Tenure – Percentage of households that are social rented in OA	Numerical	Census, 2011	A, B, C
Economic Activity – Percentage of economically active people in OA	Numerical	Census, 2011	A, B, C
Percentage of households in OA where English is first language of all adults	Numerical	Census, 2011	A, B, C
Percentage of single-person households in OA	Numerical	Census, 2011	A, B, C
Percentage of lone-parent households in OA	Numerical	Census, 2011	A, B, C
Percentage of people in OA with 'bad' or 'very bad' general health	Numerical	Census, 2011	A
Percentage of people in OA with limited long-term health	Numerical	Census, 2011	A, B, C
Police – Anti-social behaviour – Percentage of crimes in the LSOA (as a percentage of total number of anti-social behaviour crimes in whole city), 2011	Numerical	Police database (Home Office, 2016)	A, B, C
Police – Burglary – Percentage of crimes in the LSOA (as a percentage of total number of burglaries in whole city), 2011	Numerical	Police database (Home Office, 2016)	A, B, C
Police – Violent crime – Percentage of crimes in the LSOA (as a percentage of total number of violent crimes in whole city), 2011	Numerical	Police database (Home Office, 2016)	A, B, C

B2: Set 1 Results for Predicting planned/unplanned endings

Results from the Set 1 models detailed in section 7.3.7.1. The models predicted planned/unplanned endings with/without further treatment for each family. The target attribute was therefore a categorical attribute with four possible values (planned ending with further interventions, planned ending with no further interventions, unplanned ending with further interventions, unplanned ending with no further interventions). For each method (decision tree,

random forest, generalized boosted models, and logistic regression) a model was built that utilised all the data, and also separate models for each cluster. For each of these, the three different datasets (A, B, C) were utilised where appropriate (only A and B for the cluster-level models). The results of these models, in terms of their accuracy on the test dataset, together with any model parameters are contained in the following section.

Decision Tree models

Built utilising the 'rpart' R package, with 10-fold cross-validation on the training dataset. A complete tree was built (CP=0), which was then pruned to the lowest cross-validated error rate. Models were built utilising each of the three datasets (A, B, C). The performance of each pruned model was tested on the test dataset. The best model performance, which dataset this came from and the CP value, is listed in the table. In some cases, each dataset had the same result, so all are listed. For some clusters the algorithm could not find a useful model, this is denoted by a dash. The size of the test dataset is listed in the first column.

	Baseline accuracy	Test set accuracy	Dataset	CP value
All data (n=611)	47.1%	46.5%	A, B, C	0.0075, 0.0067, 0.00991
Cluster 1 (n=82)	48.8%	41.5%	A	0.024
Cluster 2 (n=97)	44.3%	38.1%	B	0.032
Cluster 3 (n=34)	41.2%	**44.1%**	B	0.089
Cluster 8 (n=62)	45.2%	-	-	-
Cluster 9 (n=68)	45.6%	-	-	-
Cluster 10 (n=52)	38.5%	**46.2%**	B	0.055
Cluster 11 (n=172)	54.7%	-	-	-

The models predicting outcome for clusters 3 and 10 had an improvement over baseline accuracy, and as such might be considered an improvement over guessing. For each of these, the variable importance scores are listed:

Cluster 3: Dataset B produced the best performance. The model indicated that the type of intervention a family received was most the important factor overall. More specifically, there was one split in the tree: which was whether a family received FF intervention treatment. If they did, a planned ending with further interventions was predicted, if not a planned ending with no further interventions was predicted. The model did not predict unplanned endings (with/without further interventions), and so despite having a small improvement over baseline accuracy may not be very useful.

Attribute	Variable importance score
Intervention type	48
Percentage of people with no qualifications in OA	12
Religion – Percentage of people who are Christian in OA	12
Number of people in the family	11
Economic Activity – Percentage of economically active people in OA	9
Percentage of households in OA not deprived	8

Cluster 10: Dataset B had the best performance. The model indicated that the housing situations in the area where a family lived (percentage of households that were owned and socially rented, together with the prevalence of lone-parent households) was the most important factor overall. However, despite having a decent improvement over the baseline accuracy (just under 8%), the model only predicted planned endings (with/without further interventions).

Attributes	Variable importance score
Tenure – Percentage of households that are owned in OA	33
Tenure – Percentage of households that are social rented in OA	26
Percentage of lone-parent households in OA	23
Percentage of people with no qualifications in OA	6
Percentage of households in OA not deprived	5
Ethnic Group – Percentage 'white' people in OA	4
Percentage of unauthorised school absence for the previous year, for the family	1
Percentage of people born in the UK in OA	1

Random Forest models

Built using the 'randomForest' R package, with 10-fold cross-validation on the training dataset. A forest with 1000 trees was built, and then the optimal number of trees selected using the cross-validated error rate. Models were built utilising each of the three datasets (A, B, C). For each the optimal model was tested on the test dataset. Any records with missing values were removed from the training dataset, as the particular algorithm cannot deal with them (this accounted for only 4 records). The best model performance (in terms of accuracy on the test dataset), which dataset this came from and the number of trees, together with the number of records in the test dataset, is listed below:

	Baseline accuracy	Test set accuracy	Dataset	Number of trees
All data (n=608)	47.1%	43.6%	B	175
Cluster 1 (n=82)	48.8%	39.0%	B	100
Cluster 2 (n=97)	44.3%	43.3%	B	25
Cluster 3 (n=34)	41.2%	41.2%	A	240
Cluster 8 (n=62)	45.2%	**48.4%**	A	145
Cluster 9	45.6%	44.1%	A	100

(n=68)				
Cluster 10 (n=52)	38.5%	**48.1%**	B	30
Cluster 11 (n=172)	54.7%	50.0%	B	300

The models predicting outcome for clusters 8 and 10 had an improvement over baseline accuracy. For each of these, the variable importance scores, in terms of mean decrease in accuracy, are listed:

Cluster 8: This utilised dataset A and indicated that, aside from the number of address changes a family had, it was mostly place-based data that had higher importance to the model. The model accuracy was only a small improvement over baseline (just over 3%), however, unlike many of the other models, it did predict 3 unplanned endings (without further interventions), together with planned endings (with/without further intervention).

Attribute	Variable importance score
Percentage of people born in the Europe in OA	2.46
Number of changes of address for the previous year, for the family	1.69
Ethnic Group – Percentage 'Other' people in OA	1.49
Ethnic Group – Percentage 'Black/African/Caribbean/Black British' people in OA	1.49
Percentage of households in OA where English is first language of all adults	1.25
Religion – Percentage of people who are Christian in OA	1.09
Percentage of unauthorised absence for the previous year, for the family	1.09
Police – Burglary – Percentage of crimes in the LSOA (as a percentage of total number of burglaries in whole city), 2011	1.02
Percentage of people born in the UK in OA	0.61
Ethnic Group – Percentage 'Asian/Asian British' people in OA	0.57
Police – Anti-social behaviour – Percentage of crimes in the LSOA (as a percentage of total number of anti-social behaviour crimes in whole city), 2011	0.55
Percentage of people in OA with limited long-term health	0.53
Percentage of people born in the Middle East and Asia in OA	0.48
Number of females in the family	0.39
Percentage of people in OA with 'bad' or 'very bad' general health	0.39
Number of people in the family	0.34
Number of children in the family	0.33
Tenure – Percentage of households that are privately rented in OA	0.30
Religion – Percentage of people who are Muslim in OA	0.22
Percentage of households in OA not deprived	0.10
Percentage of people born in the Americas and Caribbean in OA	0.03

Cluster 10: This utilised dataset A and indicated that it was mostly place-based data that had higher importance to the model. The model accuracy was a decent improvement over baseline (just under 10%); it did predict 1 unplanned ending (without further interventions), but in general predicted only planned endings (with/without further intervention).

Attribute	Variable importance score
Percentage of lone-parent households in OA	2.15
Percentage of people born in the UK in OA	2.14

Tenure – Percentage of households that are social rented in OA	2.10
Economic Activity – Percentage of economically active people in OA	1.73
Percentage of households in OA not deprived	1.54
Percentage of people in OA with limited long-term health	1.51
Percentage of unauthorised absence for the previous year, for the family	1.34
Religion – Percentage of people who are Christian in OA	0.97
Ethnic Group – Percentage 'white' people in OA	0.96
Tenure – Percentage of households that are owned in OA	0.83
Police – Anti-social behaviour – Percentage of crimes in the LSOA (as a percentage of total number of anti-social behaviour crimes in whole city), 2011	0.61
Number of children in the family	0.26
Percentage of people with no qualifications in OA	0.14
Intervention type	0.12
Percentage of households in OA where English is first language of all adults	0.09

Generalized Boosted models

Built utilising the 'gbm' R package, with 10-fold cross-validation on the training dataset. Each model built had 3000 trees; the best iteration of these was selected using the lowest cross-validated error rate. Models were built utilising each of the three datasets. For each the performance of the optimal model was tested on the test dataset. The best model performance, which dataset this came from and the best iteration, is listed in the table. The table also details the size of the test dataset.

	Baseline accuracy	Test set accuracy	Dataset	Number of trees
All data (n=611)	47.1%	46.8%	B	32
Cluster 1 (n=82)	48.8%	41.5%	B	12
Cluster 2 (n=97)	44.3%	**48.5%**	B	22
Cluster 3 (n=34)	41.2%	**50.0%**	B	16
Cluster 8 (n=62)	45.2%	43.6%	A	12
Cluster 9 (n=68)	45.6%	**48.5%**	B	14
Cluster 10 (n=52)	38.5%	**50.0%**	B	14
Cluster 11 (n=172)	54.7%	54.7%	B	20

The models predicting outcome for clusters 2, 3, 9 and 10 had performance on the test dataset that was better than the baseline accuracy. For each of these the variable importance scores are listed:

Cluster 2: This utilised dataset B and indicated that the type of intervention a family received was most important, followed by levels of crime in the area that a family lived. The model accuracy was a small improvement over baseline (just over 4%); it predicted mostly planned endings (with/without further interventions) but did pick up on one unplanned ending (without further interventions).

Attribute	Variable importance score
Intervention type	23.98
Police – Anti-social behaviour – Percentage of crimes in the LSOA (as a percentage of total number of anti-social behaviour crimes in whole city), 2011	17.54
Police – Burglary – Percentage of crimes in the LSOA (as a percentage of total number of burglaries in whole city), 2011	14.06
Police – Violent crime – Percentage of crimes in the LSOA (as a percentage of total number of violent crimes in whole city), 2011	9.12
Percentage of people born in the UK in OA	8.54
Percentage of unauthorised absence for the previous year, for the family	6.85
Percentage of households in OA where English is first language of all adults	5.84
Economic Activity – Percentage of economically active people in OA	5.17
Tenure – Percentage of households that are owned in OA	4.46
Percentage of households in OA not deprived	1.27
Ethnic Group – Percentage 'white' people in OA	1.13
Percentage of single-person households in OA	1.06
Number of adults in the family	0.98

Cluster 3: This utilised dataset B and indicated that the first intervention type was most important, followed by mostly place-based data. The model accuracy had a good improvement over the baseline accuracy (just under 9%); it predicted mostly planned endings (with/without further interventions) but did pick up two unplanned endings (without further interventions).

Attribute	Variable importance score
Intervention type	46.82
Percentage of single-person households in OA	13.57
Police – Anti-social behaviour – Percentage of crimes in the LSOA (as a percentage of total number of anti-social behaviour crimes in whole city), 2011	7.35
Religion – Percentage of people who are Christian in OA	7.32
Percentage of households in OA not deprived	6.57
Police – Violent crime – Percentage of crimes in the LSOA (as a percentage of total number of violent crimes in whole city), 2011	6.57
Percentage of people with no qualifications in OA	5.87
Ethnic Group – Percentage 'white' people in OA	3.43
Family had one or more CPPs in the previous year	2.52

Cluster 9: This utilised dataset B and used only place-based data in the model (nothing pertaining directly to the family itself had importance). The model accuracy was a small improvement over

baseline (just under 3%); however, it only predicted planned endings (with/without further interventions).

Attribute	Variable importance score
Tenure – Percentage of households that are owned in OA	21.37
Religion – Percentage of people who are Christian in OA	13.95
Economic Activity – Percentage of economically active people in OA	13.94
Police – Anti-social behaviour – Percentage of crimes in the LSOA (as a percentage of total number of anti-social behaviour crimes in whole city), 2011	13.54
Police – Burglary – Percentage of crimes in the LSOA (as a percentage of total number of burglaries in whole city), 2011	12.57
Tenure – Percentage of households that are social rented in OA	12.54
Percentage of people with no qualifications in OA	10.50
Percentage of lone-parent households in OA	1.59

Cluster 10: This utilised dataset B and identified the intervention type as most important, followed by place-based data. The model accuracy was a good improvement over baseline (just under 11%); even so, the model mostly predicted planned endings (with/without further interventions), and only detected one unplanned ending (without further interventions).

Attribute	Variable importance score
Intervention type	19.33
Percentage of single-person households in OA	19.24
Percentage of households in OA not deprived	14.42
Police – Violent crime – Percentage of crimes in the LSOA (as a percentage of total number of violent crimes in whole city), 2011	11.40
Tenure – Percentage of households that are social rented in OA	8.70
Tenure – Percentage of households that are owned in OA	7.85
Percentage of households in OA where English is first language of all adults	7.43
Police – Burglary – Percentage of crimes in the LSOA (as a percentage of total number of burglaries in whole city), 2011	5.07
Police – Anti-social behaviour – Percentage of crimes in the LSOA (as a percentage of total number of anti-social behaviour crimes in whole city), 2011	3.61
Religion – Percentage of people who are Christian in OA	2.94

Logistic Regression models

Built using the 'nnet' R package. Multinomial logistic regression was performed, as this is suitable for a categorical target attribute. The data was scaled to between zero and one and records with missing values removed. It should be noted that the models were misspecified as some of the predictors were correlated; and generally, it might not be helpful to use so many predictors with a regression model, but for a direct comparison the same dataset was utilised for all methods. Models were built utilising each of the three datasets (A, B, C). Each model was tested on the test dataset. The best model performance, in terms of test set accuracy, which dataset this came from and the Aikaike Information Criterion, is listed the table. In some cases, each dataset had the same test set accuracy, so all are listed. The size of the test dataset is included in the first column.

	Baseline accuracy	Test set accuracy	Dataset	AIC
All data (n=611)	47.1%	45.7%	B	3272.7
Cluster 1 (n=86)	47.7%	44.2%	B	476.9
Cluster 2 (n=99)	42.4%	42.4%	A	590.8
Cluster 3 (n=33)	48.5%	48.5%	A	294.0
Cluster 8 (n=61)	41.9%	31.2%	A, B	433.0, 385.1
Cluster 9 (n=64)	46.9%	29.7%	A	527.1
Cluster 10 (n=43)	37.2%	32.6%	A	412.3
Cluster 11 (n=180)	56.1%	51.1%	B	907.1

Overall, none of the logistic regression models could beat the baseline accuracy, however a few matched it. This meant that none of the models were considered useful.

B3: Set 2: Results for Predicting 'improvement'

Results from the Set 2 models detailed in section 7.3.7.2. The models predicted whether or not a family would have an 'improvement'. That is, whether the events that they had in the year prior to intervention would have stopped or decreased in the year following the start of intervention. The target attribute was therefore a dichotomous attribute with two possible values (improvement, or no improvement).

For each method (decision tree, random forest, generalized boosted models, and logistic regression) a model was built that utilised all the data, and also separate models for each cluster. For each of these, the three different datasets (A, B, C) were utilised where appropriate (only A and B for the cluster-level models). The results of these models, in terms of their accuracy on the test dataset, together with any model parameters are contained in the following tables.

Decision Tree models

Built utilising the 'rpart' R package, with 10-fold cross-validation on the training dataset. A complete tree was built (CP=0), which was then pruned to the lowest cross-validated error rate. Models were built utilising each of the three datasets (A, B, C). The performance of each pruned model was tested on the test dataset. The best model performance, which dataset this came from and the CP value, is listed in the table. In some cases, each dataset had the same result, so all are listed. For some clusters the algorithm could not find a useful model, this is denoted by a dash. The size of the test dataset is listed in the first column.

	Baseline accuracy	Test set accuracy	Dataset	CP value
All data (n=500)	54.2%	**67.0%**	A, B, C	0.017, 0.0067, 0.015
Cluster 1 (n=69)	75.4%	53.6%	A	0.02
Cluster 2 (n=75)	57.3%	-	-	-
Cluster 3 (n=28)	71.4%	67.9%	A	0.013
Cluster 8 (n=49)	65.3%	53.1%	B	0.06
Cluster 9 (n=57)	54.4%	40.4%	B	0.009
Cluster 10 (n=44)	68.2%	-	-	-
Cluster 11 (n=142)	72.5%	**73.2%**	B	0.0

The models that had a test set accuracy greater than the baseline were those which utilised the whole dataset and the cluster 11 model. The details for each are:

All data: the accuracy was the same for each set of attributes used (A, B, C). No matter the set of attributes, the tree produced was the same (although there was a little variation in the surrogate values). The tree had only one split, which was whether a family had children or not. If they did not have children, improvement was predicted; if they did, no improvement was predicted. The variable importance scores are listed below:

Set A		Set B		Set C	
Attribute	**Score**	**Attribute**	**Score**	**Attribute**	**Score**
Number of children in family	66	Number of children in family	66	Number of children in family	60
Number of people in family	34	Number of people in family	34	Number of people in family	31
				Cluster assignment	9

Cluster 11: this utilised dataset B and had a very small improvement over baseline accuracy (less than 1%). The most important attributes pertained to family size; how many children there were in the family, and how many people overall. The variable importance scores are listed below:

Attribute	Variable importance score
Number of children in the family	21
Number of people in the family	15
Religion – Percentage of people who are Christian in OA	6
Tenure – Percentage of households that are owned in OA	6
Number of adults in the family	5
Intervention type	5
Percentage of households in OA not deprived	5
Tenure – Percentage of households that are social rented in OA	5
Percentage of people with no qualifications in OA	5
Percentage of lone-parent households in OA	4
Percentage of single-person households in OA	4

Police – Anti-social behaviour – Percentage of crimes in the LSOA (as a percentage of total number of anti-social behaviour crimes in whole city), 2011	4
Economic Activity – Percentage of economically active people in OA	3
Police – Burglary – Percentage of crimes in the LSOA (as a percentage of total number of burglaries in whole city), 2011	3
Percentage of households in OA where English is first language of all adults	3
Police – Violent crime – Percentage of crimes in the LSOA (as a percentage of total number of violent crimes in whole city), 2011	3
Ethnic Group – Percentage 'white' people in OA	2
Percentage of people born in the UK in OA	1
Percentage of people in OA with limited long-term health	1

Random Forest models

Built using the 'randomForest' R package, with 10-fold cross-validation on the training dataset. A forest with 1000 trees was built, and then the optimal number of trees selected using the cross-validated error rate. Models were built utilising each of the three datasets (A, B, C). For each the optimal model was tested on the test dataset. Any records with missing values were removed from the training dataset, as the particular algorithm cannot deal with them (this accounted for only 4 records). The best model performance (in terms of accuracy on the test dataset), which dataset this came from and the number of trees, together with the number of records in the test dataset, is listed in the table:

	Baseline accuracy	Test set accuracy	Dataset	Number of trees
All data (n=500)	54.4%	**62.5%**	A, C	145,230
Cluster 1 (n=69)	75.4%	**76.8%**	A	80
Cluster 2 (n=75)	57.3%	57.3%	B	80
Cluster 3 (n=28)	71.4%	67.9%	B	40
Cluster 8 (n=49)	65.3%	61.2%	A	50
Cluster 9 (n=57)	54.4%	50.9%	B	180
Cluster 10 (n=44)	68.2%	65.9%	A	140
Cluster 11 (n=142)	72.5%	**74.3%**	B	50

The models predicting outcome for all the data, and cluster 1, had an improvement over baseline accuracy. For each of these, the variable importance scores, in terms of mean decrease in accuracy, are listed:

All data: The models utilising datasets A and C both had the same prediction accuracy on the test dataset, although they produced different models. For each, the model accuracy was an improvement over baseline accuracy of just over 8%. The number of children in the family, and

number of people overall, had high importance to both models. The cluster assignment was most important to the set C model, whereas the attributes that contribute to cluster assignment (pertaining to events) were more important to the set A model (as this did not have the cluster assignment attribute). The full list of variable importance scores, for sets A and C are listed:

Set A		Set C	
Attribute	**Variable importance score**	**Attribute**	**Variable importance score**
Number of children in the family	13.89	Cluster assignment	24.50
Percentage of unauthorised absence for the previous year, for the family	9.02	Number of children in the family	21.44
Number of people in the family	8.20	Number of people in the family	15.07
Number of males in the family	6.86	Percentage of single-person households in OA	4.79
Family had one or more CIN events in the previous year	5.58	Intervention type	4.52
Number of females in the family	5.43	Economic Activity – Percentage of economically active people in OA	4.17
Number of criminal offences committed by children for the previous year, for the family	3.88	Tenure – Percentage of households that are owned in OA	3.58
Percentage of people in OA with limited long-term health	3.66	Percentage of people born in the UK in OA	3.56
Family had one or more CPPs in the previous year	3.63	Percentage of people in OA with limited long-term health	3.22
Number of school exclusions for the previous year, for the family	3.43	Ethnic Group – Percentage 'white' people in OA	3.15
Percentage of people born in the UK in OA	2.22	Tenure – Percentage of households that are social rented in OA	2.74
Intervention type	1.87	Percentage of people with no qualifications in OA	2.15
Economic Activity – Percentage of economically active people in OA	1.77	Percentage of households in OA not deprived	1.98
Family had one or more events classed as domestic abuse in the previous year	1.77	Percentage of lone-parent households in OA	1.86
Religion – Percentage of people who are Muslim in OA	1.50	Percentage of households in OA where English is first language of all adults	1.21
Number of criminal offences committed by adults for the previous year, for the family	1.50	Religion – Percentage of people who are Christian in OA	0.97
Ethnic Group – Percentage 'Black/African/Caribbean/Black British' people in OA	1.46	Police – Violent crime – Percentage of crimes in the LSOA (as a percentage of total number of violent crimes in whole city), 2011	0.67
Tenure – Percentage of households that are owned in OA	1.46	Number of adults in the family	0.31
Tenure – Percentage of households that are social rented in OA	1.39		
Percentage of households in OA where English is first language of all adults	1.35		

Percentage of people with no qualifications in OA	1.32		
Number of changes of address for the previous year, for the family	1.30		
Ethnic Group – Percentage 'white' people in OA	1.27		
Family had one or more LAC events in the previous year	0.95		
Percentage of people in OA with 'bad' or 'very bad' general health	0.85		
Religion – Percentage of people who have no religion in OA	0.84		
Police – Violent crime – Percentage of crimes in the LSOA (as a percentage of total number of violent crimes in whole city), 2011	0.46		
Percentage of people born in the Middle East and Asia in OA	0.37		
Percentage of people born in the Africa in OA	0.16		
Percentage of lone-parent households in OA	0.12		

Cluster 1: This utilised dataset A and indicated that the number of children in the family was most important. The model accuracy was only a small improvement over baseline (just over 1%). The variable importance scores were:

Attribute	Variable importance score
Number of children in the family	2.89
Percentage of households in OA where English is first language of all adults	2.33
Percentage of single-person households in OA	2.16
Percentage of people born in the UK in OA	2.09
Police – Anti-social behaviour – Percentage of crimes in the LSOA (as a percentage of total number of anti-social behaviour crimes in whole city), 2011	2.09
Ethnic Group – Percentage 'Other' people in OA	1.45
Number of females in the family	1.44
Tenure – Percentage of households that are social rented in OA	1.42
Number of males in the family	1.41
Number of criminal offences committed by children for the previous year, for the family	1.23
Percentage of people in OA with limited long-term health	1.14
Percentage of people born in the Africa in OA	1.05
Percentage of lone-parent households in OA	0.84
Number of adults in the family	0.71
Number of people in the family	0.47
Number of school exclusions for the previous year, for the family	0.39
Religion – Percentage of people who are Muslim in OA	0.36
Religion – Percentage of people who are Christian in OA	0.36
Percentage of people born in the Antarctica and Oceania in OA	0.26
Tenure – Percentage of households that are privately rented in OA	0.16

Cluster 11: This utilised dataset B and indicated that attributes pertaining to family size were most important, together with place-based attributes. The model accuracy was a small improvement over baseline (just under 2%).

Attribute	Variable importance score
Number of children in the family	7.11
Number of people in the family	4.84
Police – Anti-social behaviour – Percentage of crimes in the LSOA (as a percentage of total number of anti-social behaviour crimes in whole city), 2011	3.94
Number of adults in the family	3.33
Tenure – Percentage of households that are owned in OA	1.92
Percentage of single-person households in OA	1.65
Economic Activity – Percentage of economically active people in OA	1.43
Intervention type	0.95
Percentage of people with no qualifications in OA	0.78
Percentage of households in OA where English is first language of all adults	0.66
Religion – Percentage of people who are Christian in OA	0.58
Percentage of people in OA with limited long-term health	0.53
Tenure – Percentage of households that are social rented in OA	0.51
Percentage of lone-parent households in OA	0.26
Ethnic Group – Percentage 'white' people in OA	0.13

Generalized Boosted models

Built utilising the 'gbm' R package, with 10-fold cross-validation on the training dataset. Each model built had 3000 trees; the best iteration of these was selected using the lowest cross-validated error rate. Models were built utilising each of the three datasets. For each the performance of the optimal model was tested on the test dataset. The best model performance, which dataset this came from and the best iteration, is listed in the table. The table also details the size of the test dataset.

	Baseline accuracy	Test set accuracy	Dataset	Number of trees
All data (n=611)	54.2%	**65.8%**	A	17
Cluster 1 (n=82)	75.4%	75.4%	A	17
Cluster 2 (n=97)	57.3%	**64.0%**	A	5
Cluster 3 (n=34)	71.4%	71.4%	A	8
Cluster 8 (n=62)	65.3%	51.0%	A, B	18, 11
Cluster 9 (n=68)	54.4%	49.1%	B	1
Cluster 10 (n=52)	68.2%	63.6%	A	9
Cluster 11 (n=172)	72.5%	**76.1%**	A	22

The models predicting outcome for all the data, and clusters 2 and 11 had performance on the test dataset that was better than the baseline accuracy. For each of these the variable importance scores are listed:

All data: This utilised dataset A and only four attributes had any importance, with the number of children in the family being most important. The model accuracy was a good improvement over baseline (just over 11%).

Attribute	Variable importance score
Number of children in the family	49.28
Number of people in the family	26.92
Percentage of unauthorised absence for the previous year, for the family	18.98
Intervention type	4.82

Cluster 2: This utilised dataset A and only three attributes had any importance, with the percentage of lone-parent households in the area that a family lived being most important, followed by the percentage of privately rented households. The model accuracy was a decent improvement over baseline (just under 7%).

Attribute	Variable importance score
Percentage of lone-parent households in OA	41.42
Tenure – Percentage of households that are privately rented in OA	36.36
Number of females in the family	22.22

Cluster 11: This utilised dataset A and only six attributes had any importance, with the number of children, and people overall, in the family being most important. The model accuracy was a small improvement over baseline (just under 4%).

Attribute	Variable importance score
Number of children in the family	63.24
Number of people in the family	22.63
Percentage of people born in the Middle East and Asia in OA	4.52
Intervention type	3.57
Tenure – Percentage of households that are owned in OA	3.32
Police – Violent crime – Percentage of crimes in the LSOA (as a percentage of total number of violent crimes in whole city), 2011	2.73

Logistic Regression models

Built using the 'glm' R package. The data was scaled to between zero and one and records with missing values removed. It should be noted that the models were misspecified as some of the predictors were correlated; and generally, it might not be helpful to use so many predictors with a regression model, but for a direct comparison the same dataset was utilised for all methods. Models were built utilising each of the three datasets (A, B, C). Each model was tested on the test dataset. The best model performance, in terms of test set accuracy, which dataset this came from and the Aikaike Information Criterion, is listed in the table. In some cases, each dataset had the same test set accuracy, so all are listed. The size of the test dataset is included in the first column.

	Baseline accuracy	Test set accuracy	Dataset	AIC
All data	54.2%	**64.0%**	C	1375.5

(n=500)				
Cluster 1 (n=67)	68.7%	68.7%	A	203.5
Cluster 2 (n=81)	53.1%	51.9%	A	248.68
Cluster 3 (n=32)	81.3%	68.8%	B	60
Cluster 8 (n=41)	51.2%	48.8%	A, B	145
Cluster 9 (n=54)	63.0%	61.1%	A	222.19
Cluster 10 (n=45)	68.9%	64.4%	B	134.23
Cluster 11 (n=140)	68.6%	**70.0%**	B	333.74

Two models had accuracy on the test dataset that beat the baseline accuracy, that which utilised all the data, and the cluster 11 model.

All data: This utilised dataset C and the model accuracy on the test dataset was a decent improvement over baseline (just over 8%). 'No improvement' was set as the reference level. The model had an AIC value of 1375.5 and residual deviance of 1309.5 on 1129 degrees of freedom. The full summary is listed below, with values that were significant at the $p < 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***) indicated. The model indicated that cluster assignment (particularly, cluster 11), the percentage of single person households in the area the family lived in, and receiving the FIP intervention type were all important to the model.

Coefficient:	Estimate	Standard Error	z value	Pr (>|z|)
(Intercept)	-0.32024	1.55462	-0.206	0.836797
Cluster assignment: 2	0.89385	0.26243	3.406	0.000659 ***
Cluster assignment: 3	0.54565	0.35758	1.526	0.127024
Cluster assignment: 4	1.48294	0.48656	3.048	0.002305 **
Cluster assignment: 5	1.18555	0.75428	1.572	0.116002
Cluster assignment: 6	1.64708	0.43780	3.762	0.000168 ***
Cluster assignment: 7	-0.47868	0.80029	-0.598	0.549749
Cluster assignment: 8	0.24432	0.29083	0.840	0.400869
Cluster assignment: 9	1.13736	0.26968	4.218	0.0000247 ***
Cluster assignment: 10	0.58091	0.30703	1.892	0.058490
Cluster assignment: 11	1.78834	0.24475	7.307	0.0000000000003 ***
Ethnic Group – Percentage 'white' people in OA	-0.05694	1.16972	-0.049	0.961114
Percentage of households in OA not deprived	-2.72488	1.51775	-1.795	0.072600
Percentage of people born in the UK in OA	-1.76055	1.99029	-0.885	0.376390
Percentage of people with no qualifications in OA	0.87637	1.54889	0.566	0.571527
Religion – Percentage of people who are Christian in OA	-0.89944	1.10451	-0.814	0.415454
Tenure – Percentage of households that are owned in OA	-0.94746	0.84980	-1.115	0.264884
Tenure – Percentage of households that are social rented in OA	-0.23472	0.64452	-0.364	0.715726
Economic Activity – Percentage of economically active people in OA	2.14631	1.43960	1.491	0.135987
Percentage of households in OA where English is first language of all adults	3.39647	1.79956	.1887	0.059108
Percentage of single-person households in OA	-2.73654	0.97384	-2.810	0.004954 **

Percentage of lone-parent households in OA	-4.09520	1.78497	-2.294	0.021775 *
Percentage of people in OA with limited long-term health	-0.69836	2.21664	-0.315	0.752722
Police – Anti-social behaviour – Percentage of crimes in the LSOA (as percent of total in whole city), 2011	1.06963	1.14781	0.932	0.351395
Police – Violent crime – Percentage of crimes in the LSOA (as percent of total in whole city), 2011	0.28571	2.02185	0.141	0.887626
Police – Burglary – Percentage of crimes in the LSOA (as percent of total in whole city), 2011	1.91246	1.92456	0.994	0.320364
Intervention type: CFPT	0.05811	0.18685	0.322	0.755791
Intervention type: FF	-0.23409	0.23353	-1.002	0.316150
Intervention type: FINIS	0.42468	0.46740	0.887	0.374970
Intervention type: FIP	-0.53533	0.19668	-2.722	0.006492 **
Number of people in the family	-8.70434	6.43471	-1.353	0.176146
Number of adults in the family	3.76393	3.54724	1.061	0.288650
Number of children in the family	3.11680	5.24948	0.594	0.522690

Cluster 11: This utilised dataset B and the model accuracy on the test dataset was a very small improvement over baseline (just over 1%). 'Improvement' was set as the reference level. The model had an AIC value of 333.74 and residual deviance of 287.74 on 303 degrees of freedom. The full summary is listed below; no values were significant at the $p < 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***) level. Given that the model was only marginally better than the baseline accuracy, it may not be very useful.

Coefficient:	**Estimate**	**Standard Error**	**z value**	**Pr (>|z|)**
(Intercept)	3.0707	3.6030	0.852	0.3941
Ethnic Group – Percentage 'white' people in OA	2.1356	2.7475	0.777	0.4370
Percentage of households in OA not deprived	1.5040	3.7954	0.396	0.6919
Percentage of people born in the UK in OA	-3.0991	4.9112	-0.631	0.5280
Percentage of people with no qualifications in OA	1.8585	3.3652	0.552	0.5808
Religion – Percentage of people who are Christian in OA	-1.5133	2.6328	-0.575	0.5653
Tenure – Percentage of households that are owned in OA	-0.8717	2.0422	-0.427	0.6695
Tenure – Percentage of households that are social rented in OA	-0.6239	1.5177	-0.411	0.6810
Economic Activity – Percentage of economically active people in OA	3.3183	3.4198	0.970	0.319
Percentage of households in OA where English is first language of all adults	-0.6450	4.3058	-0.150	0.8809
Percentage of single-person households in OA	-1.6283	2.3282	-0.699	0.4842
Percentage of lone-parent households in OA	1.0732	4.0815	0.263	0.7926
Percentage of people in OA with limited long-term health	6.0825	5.8698	1.031	0.3023
Police – Anti-social behaviour – Percentage of crimes in the LSOA (as percent of total in whole city), 2011	-0.9278	2.5197	-0.368	0.7127
Police – Violent crime – Percentage of crimes in the LSOA (as percent of total in whole city), 2011	-5.6321	4.5422	-1.240	0.2150
Police – Burglary – Percentage of crimes in the LSOA (as percent of total in whole city), 2011	-1.9216	4.1982	00.458	0.6472
Intervention type: CFPT	0.5794	0.4153	1.395	0.1630
Intervention type: FF	-0.5660	0.5138	-1.102	0.2706
Intervention type: FINIS	-0.5619	1.2940	-0.434	0.6641
Intervention type: FIP	-0.7098	0.3995	-1.777	0.0756
Number of people in the family	-182.4985	9710.1787	-0.019	0.9850
Number of adults in the family	96.5253	5296.4614	0.018	0.9855
Number of children in the family	138.8173	7944.6917	0.017	0.9861

B4: Predicting 'improvement' for families with and without children

Results from the models detailed in section 7.3.7.3. The data was split into families with children (n=1372) and without children (n=296), and the models predicted whether or not a family would have an 'improvement'. That is, whether the events that they had in the year prior to intervention would have stopped or decreased in the year following the start of intervention. The target attribute was therefore a dichotomous attribute with two possible values (improvement, or no improvement). As with the models built in the previous sections, each dataset was split into a training and testing dataset (70:30 split) and the machine learning models utilised 10-fold cross-validation on the training dataset to determine the optimal model which was then tested on the test dataset. The logistic regression model was built on the training set and tested on the test set.

For each method (decision tree, random forest, generalized boosted models, and logistic regression) an overall model was built that utilised all the data; there were no cluster-level models. For each of these, the three different datasets (A, B, C) were utilised. The results of these models, in terms of their accuracy on the test dataset are contained in the table:

	Baseline accuracy	Test set accuracy	Dataset
Families with children:			
Decision tree	63.1%	62.1%	C
Random forest	63.1%	62.1%	B
Generalized boosted model	63.1%	62.9%	B, C
Logistic regression	63.1%	**64.1%**	A
Families without children:			
Decision tree	87.5%	84.1%	C
Random forest	87.5%	86.4%	B, C
Generalized boosted model	87.5%	**88.6%**	C
Logistic regression	87.5%	85.2%	B

Whilst a couple of the models had a marginal improvement over the baseline accuracy (around 1%) it would seem that the models were not very useful and could not really beat the accuracy attained simply from guessing.